

DOCTORAL DISSERTATION SERIES

TITLE THE CONSTRUCTION AND VALIDATION OF A  
TEST TO MEASURE SOME OF THE  
INDUCTIVE ASPECTS OF SCIENTIFIC  
THINKING

AUTHOR MARY ALICE HORSWILL BURMESTER

UNIVERSITY MICH STATE COLL. DATE 1951

DEGREE Ed. D. PUBLICATION NO. 4713



UNIVERSITY MICROFILMS

ANN ARBOR • MICHIGAN

COPYRIGHTED  
BY  
Mary Alice Horswill Burmester  
1953

THE CONSTRUCTION AND VALIDATION OF A TEST  
TO MEASURE SOME OF THE INDUCTIVE  
ASPECTS OF SCIENTIFIC THINKING

BY

Mary Alice Burmester

\*  
\* \*  
\*

A THESIS

Submitted to the Graduate School of Michigan  
State College of Agriculture and Applied  
Science in partial fulfillment of the  
requirements for the degree of

DOCTOR OF EDUCATION

Department of Education

1 9 5 1

## ACKNOWLEDGMENTS

The writer wishes to express appreciation for the assistance given by Dr. Victor H. Noll, thesis adviser, and to the other members of the advisory committee. She also wishes to thank Dr. Clarence H. Nelson of the Board of Examiners of Michigan State College for valuable suggestions on the construction of test items and to acknowledge the cooperation of all of the members of the Department of Biological Science of Michigan State College for their aid in the study.



THE CONSTRUCTION AND VALIDATION OF A TEST  
TO MEASURE SOME OF THE INDUCTIVE  
ASPECTS OF SCIENTIFIC THINKING

By

Mary Alice Burmester

AN ABSTRACT

Submitted to the Graduate School of Michigan  
State College of Agriculture and Applied  
Science in partial fulfillment of the  
requirements for the degree of

DOCTOR OF EDUCATION

Department of Education

1 9 5 1

Approved

*Victor H. Noll*

---

The purpose of this study was to devise a valid test to measure some of the inductive aspects of the ability to think scientifically, in the area of biological science. The educational objectives related to scientific thinking were formulated and were defined in terms of desired behaviors involved. In all, 98 behaviors were recognized as attending the critical, as opposed to the creative, aspects of scientific thinking. Nine tryout tests, consisting of a total of 637 items were constructed to evaluate these behaviors. These tests were administered during the spring term of 1950 to 168 students taking the third term of the three-term sequence of Biological Science at Michigan State College. Item validity and item difficulty were calculated for each item of the tryout tests.

Test I, The Ability to Think Scientifically, constructed from discriminating items of the tryout tests, consisted of 150 items. Test I was administered in the spring of 1950 to 500 students at the end of the three-term sequence of Biological Science, and in the fall of 1950 to another group of 240 students who had had no college biology. The reliabilities of the test for the two groups were .89 and .91 respectively. Because Test I proved too long, 25 of the poorer items, as identified by item analysis, were eliminated. The remainder constituted Test IA, The Ability to Think Scientifically. This test was administered in the fall of 1950 to 330 students who had had no college biology, and to 136 of these same students after completion of one term of Biological Science. The reliabilities for the two groups were .91 and .90 respectively.

The curricular validity of the test was established by:

1. Designing the test items to measure the behaviors involved in scientific thinking.
2. Submission of the tryout tests to competent judges for criticism.
3. Using free responses of students as items wherever feasible.
4. Careful selection of materials utilized in the construction of the test items.

Three general methods were used in the statistical validation of the test, namely,

1. Scores made on the test of the ability to think scientifically were correlated with measures of intelligence, of reading ability and of knowledge of biological facts. These correlations ranged from .33 to .51.
2. Mean scores made by students who had had no college biology were compared with mean scores made by students who had had Biological Science. The means of those having had Biological Science were significantly higher.
3. Scores made on Test IA by 143 students were compared with ratings of these students by their instructors on their ability to think scientifically. The chi-square test, a comparison of means of students receiving superior, average and inferior ratings, and a correlation of scores on the test with the ratings all gave evidence of the statistical validity of the test. The correlation between scores on the test and the ratings of the instructors was .77 for the test when administered as a pretest, and .72 when administered as a post-test.

Mary Alice Burmester  
candidate for the degree of  
Doctor of Education

Final examination, May 10, 1951, 3:00 P. M.

Dissertation: The Construction and Validation of a Test to  
Measure Some of the Inductive Aspects of  
Scientific Thinking

Outline of Studies

Major subject: Education  
Cognate area: Physiology

Biographical Items

Born, September 1, 1909, Oakland, California

Undergraduate Studies, University of California, 1926-1930

Graduate Studies, University of California, 1930-1933  
Michigan State College, 1946-1951

Experience: Teaching Assistant in Physiology, University  
of California, 1931-1934, Instructor in  
Biological Science, Michigan State College,  
1945-1948, Assistant Professor in Biological  
Science, Michigan State College, 1948-1951

Member of Kappa Delta Pi, Phi Sigma, Sigma Xi

## TABLE OF CONTENTS

CHAPTER	PAGE
I. THE BACKGROUND OF THE PROBLEM .....	1
Introduction .....	1
The problem .....	10
Statement of the problem .....	10
Delimitation of the problem .....	11
Basic assumptions of the study .....	11
Importance of the study .....	12
Organization of the remainder of the thesis	13
II. REVIEW OF RESEARCH RELATED TO THE PROBLEM .	15
Steps and skills of scientific thinking .	15
The measurement of problem-solving	
abilities .....	22
Summary concerning tests on abilities	
involved in problem-solving .....	60
Relationship between problem-solving and	
other abilities .....	63
Relation of intelligence to abilities	
involved in problem-solving .....	63
Summary of studies concerning the	
relation of intelligence to problem-	
solving .....	73
Educability in problem-solving .....	75
Summary of studies on educability in	
problem-solving .....	82

## CHAPTER

## PAGE

Relation of reading to abilities	
involved in problem-solving .....	82
Summary of studies concerning the	
relation of reading ability to	
problem-solving .....	85
Relation of factual information to the	
abilities involved in problem-solving .	85
Summary of studies concerning the rela-	
tion of knowledge of facts to problem-	
solving abilities .....	89
Summary of research related to the	
problem .....	89
III. GENERAL PROCEDURES INVOLVED IN THE DEVELOP-	
MENT OF THE TEST .....	92
IV. THE DEVELOPMENT OF THE TEST ITEMS .....	107
The formulation of the educational	
objectives .....	107
The definition of the behaviors .....	109
Methods used to determine the behaviors.	109
An outline of the behaviors .....	116
The location of the source materials from	
which the items could be constructed ..	121
The construction of the evaluation	
instruments .....	124
Analysis of the tryout tests in terms of	
the behaviors involved .....	137

## CHAPTER

## PAGE

## V. THE STATISTICAL ANALYSES OF THE TESTS AND

THE TEST ITEMS .....	144
Methods used in item-analysis .....	144
Analysis of tryout tests .....	146
Analysis of Test A - Some Steps in Scientific Thinking .....	146
Analysis of Test B - The Delimitation of Problems .....	148
Analysis of Test C - Experimental Procedures .....	150
Analysis of Test D - Organization of Data .....	152
Analysis of Test E - Evaluation of Hypothesis .....	154
Analysis of Test F - Experimentation and the Interpretation of Data .....	155
Analysis of Test G - Drawing of Conclusions .....	156
Analysis of Test H - Interpretation of Data .....	157
Analysis of Test J - Generalizations and Assumptions .....	159
Analysis of tryout tests considered as a single test .....	162
Intercorrelations of tryout test scores.	162

## CHAPTER

## PAGE

Correlations of scores on tryout tests with scores on intelligence and read- ing tests .....	171
The preparation of Test I - The Ability to Think Scientifically .....	173
Analyses of Test I and Test IA	175
Analysis of Test I - The Ability to Think Scientifically .....	175
Analysis of Test IA - The Ability to Think Scientifically .....	183
VI. THE VALIDATION OF THE TEST .....	187
The curricular validation of the test ...	188
The statistical validation of the test ..	192
Validation by correlation with measures of intelligence, reading ability, and factual information .....	192
Validation by comparison of scores of various groups .....	198
Validation by comparison of scores with ratings of students by competent judges .....	202
VII. SUMMARY AND CONCLUSIONS .....	211
Summary .....	211
Conclusions .....	219
Educational implications .....	220



## CHAPTER

## PAGE

Educational implications for Biological

Science at Michigan State College ... 220

Educational implications for science

courses in general education ..... 221

Other educational implications ..... 222

Problems suggested by the study ..... 222

LITERATURE CITED ..... 227

APPENDIX I ..... 236

APPENDIX II ..... 351

APPENDIX III ..... 383

APPENDIX IV ..... 403

# LIST OF TABLES

TABLE	PAGE
I. Behaviors Measured by the Tryout Tests ...	138
II. Pertinent Data for Test A .....	148
III. Item Analysis Data on the Seven Items of Test B which Measured Ability to Recognize Assumptions Underlying Problems .....	149
IV. Pertinent Data for Test B .....	150
V. Pertinent Data for Test C .....	151
VI. Pertinent Data for Test D .....	154
VII. Pertinent Data for Test E .....	155
VIII. Pertinent Data for Test F .....	156
IX. Pertinent Data for Test G .....	157
X. Pertinent Data for Test H .....	158
XI. Pertinent Data for Test J .....	159
XII. Comparison of Means, Standard Deviations, and Reliabilities of the Tryout Tests ..	160
XIII. Comparison of Mean Item Validities and Mean Item Difficulties of the Tryout Tests ..	161
XIV. Pertinent Data for the Tryout Test Battery.	162
XV. Intercorrelations of Tryout Test Scores ..	163
XVI. Intercorrelations of Tryout Test Scores Corrected for Attenuation .....	165
XVII. Coefficients of Determination of Tryout Tests .....	166

TABLE	PAGE
XVIII. Correlation of Total Scores on Tryout Test Battery with Each of the Tryout Tests ..	167
XIX. Multiple Correlation of Tryout Total with Two of the Tryout Tests .....	169
XX. Multiple Correlation of Tryout Tests with the Criterion - Obtained by the Wherry-Doolittle Method .....	170
XXI. Correlation of Tryout Test Scores with Intelligence Test and Reading Test Scores	172
XXII. Pertinent Data for Test I .....	177
XXIII. Comparison of Discrimination Indices and of Difficulty Indices of Identical Items as Obtained from Item Analysis of Tryout Tests and as Obtained from Item Analysis.	178
XXIV. Summary of Item Analysis Data for Tryout Test Items Used in Construction of Test I, Items of Test I, and Items of Test I Used in Construction of Test IA .....	183
XXV. Pertinent Data for Test IA .....	185
XXVI. Correlation of Tryout Test Scores and Scores on Test I with Psychological Examination Scores and Reading Test Scores	194
XXVII. Intercorrelations of Tryout Test, Psychological Examination, and Reading Test .....	195

TABLE	PAGE
XXVIII. Intercorrelation of Test I, Psychological Examination and Reading Test .....	196
XXIX. Intercorrelation of Total Tryout Test Scores and Scores on Other Tests .....	197
XXX. Comparison of Means and Standard Deviation of Test I for a Group Before Taking Biological Science with Another Group After Taking Three Terms of Biological Science	200
XXXI. Comparison of Means and Standard Deviations of Test IA on Pre-Test and Post-Test .....	201
XXXII. Expectancy Chart Showing the Comparison of Scores on Test IA Pre-Test and Ratings .	206
XXXIII. Expectancy Chart Showing the Comparison of Scores on Test IA Post-Test and Ratings	207
XXXIV. Mean Gains of Students Rated as Superior, Inferior, and Average on Test IA .....	208
XXXV. Differences in Means and Critical Ratios of Differences between Students Rated Superior and Students Rated Average and Students Rated Average and Students Rated Inferior .....	209
XXXVI. Item Analysis Data for Test A .....	246
XXXVII. Item Analysis Data for Test B .....	256
XXXVIII. Item Analysis Data for Test C .....	269

TABLE	PAGE
XXXIX. Item Analysis Data for Test D .....	279
XL. Item Analysis Data for Test E .....	289
XLI. Item Analysis Data for Test F .....	301
XLII. Item Analysis Data for Test G .....	317
XLIII. Item Analysis Data for Test H .....	340
XLIV. Item Analysis Data for Test J .....	345
XLV. Item Analysis Data for Test I .....	373

## CHAPTER I

### THE BACKGROUND OF THE PROBLEM

#### INTRODUCTION

With the growth of a general education program in the secondary schools and the lower college years there has been an increased emphasis upon the acquisition of knowledge, skills, and attitudes which are required for participation in a democratic society.<sup>1</sup> One of these skills, which has become a major objective of education, is the ability to solve problems. This objective has been stated variously by different educators. They refer to it as reflective thinking, critical thinking, clear thinking, or as scientific thinking. Although different terms are used they all refer to the kind of thinking involved in the solution of a problem.

As early as 1909, Dewey<sup>2</sup> advocated the teaching of scientific habits of mind. He asserted then and has continued to contend<sup>3</sup> that the problem of problems in our

---

<sup>1</sup> American Council on Education, Executive Committee of the Cooperative Study in General Education, Cooperation in General Education. Washington: American Council on Education. 1947. p. 12.

<sup>2</sup> John Dewey, How We Think. Boston: D. C. Heath and Company. 1909. (preface).

<sup>3</sup> John Dewey, "Method in science teaching." Science Education, 29:119-23, April, 1945.

education is to discover how to teach scientific habits of thought. Almost every major educational committee in the last twenty-five years has emphasized the importance of this instructional objective, not alone as an objective of science courses, but as an objective for general education. Evidence for this is presented in the paragraphs that follow.

Eurich,<sup>4</sup> in a report in the Thirty-eighth Yearbook of the National Society for Education said that there should be a "deepened desire to do something that will make education more effective than it has been in the past, largely, perhaps, in the hope that future generations will be able to solve better such social problems as those that baffle present-day society."

The Educational Policies Commission<sup>5</sup> in 1944 made a plea for the reorganization of the secondary schools of America. A plan was presented for the education of all American youth. The following quotation gives the broad outline of this plan:

Schools should be dedicated to the proposition that every youth in these United States - regardless of sex, economic status, geographic location, or race - should

---

<sup>4</sup> Alvin C. Eurich, "A renewed emphasis upon general education," in General Education in the American College. Thirty-eighth Yearbook of the National Society for the Study of Education, Part II, p. 6-7. Bloomington, Illinois: Public School Publishing Company, 1939.

<sup>5</sup> Educational Policies Commission, Education for All American Youth. Washington: National Education Association. 1944. p. 21.

experience a broad and balanced education which will (1) equip him to enter an occupation suited to his abilities and offering reasonable opportunity for personal growth and social usefulness; (2) prepare him to assume the full responsibilities of American citizenship; (3) give him a fair chance to exercise his right to the pursuit of happiness; (4) stimulate intellectual curiosity, engender satisfaction in intellectual achievement, and cultivate the ability to think rationally; and (5) help him to develop an appreciation of the ethical values which should undergird all life in a democratic society. It is the duty of a democratic society to provide opportunities for such education through its schools.<sup>6</sup>

Further evidence that the ability to think critically is a major objective of education is supplied by the following statement of a committee which evaluated educational objectives: "The committee believes that the ability to think reflectively and the disposition to do so in all the problem situations of life is an especially important educational objective."<sup>7</sup> This same committee stated that this ability is "peculiarly necessary in a democracy, where each is expected to take part in policy-making."<sup>8</sup>

The importance of this objective is also emphasized in the following quotation:

The responsibility of secondary schools for training citizens who can think clearly has been so long and so frequently acknowledged that it is now almost taken for granted. The educational objectives classifiable under the generic heading "clear thinking" are numerous and varied as to statement, but there can be little doubt

---

<sup>6</sup> Loc. cit.

<sup>7</sup> Progressive Education Association, Science in General Education. New York: D. Appleton-Century Company. 1938. p. 306.

<sup>8</sup> Ibid., p. 46.



concerning their fundamental importance. Although in recent years there has been increasing recognition of other responsibilities and purposes, there has been little accompanying tendency to demote clear thinking to a minor role as an educational objective. It was therefore not surprising to find considerable emphasis upon this objective in the statements of purposes submitted to the Evaluation Staff by the schools participating in the Eight-Year Study.<sup>9</sup>

The Harvard Committee<sup>10</sup> and the President's Commission on Higher Education<sup>11</sup> both recognized reflective thinking as a major objective of education. The much quoted report of the Harvard Committee on General Education stressed the values of reflective thinking. According to this report abilities which should be sought above all others in the general education program are the ability to think effectively, to communicate thought, to make relevant judgments, and to discriminate among values. The President's Commission on Higher Education included the ability "to acquire and use the skills and habits involved in critical and constructive thinking" as one of the eleven basic objectives of general education.

As may be seen from the above discussion the ability to solve problems is a stated objective of general education

---

<sup>9</sup> Eugene R. Smith, Ralph W. Tyler and the Evaluation Staff, Appraising and Recording Student Progress. New York: Harper and Brothers. 1942. p. 35.

<sup>10</sup> Harvard University, General Education in a Free Society. Cambridge: Harvard University Press. 1945. p. 65.

<sup>11</sup> President's Commission on Higher Education, Higher Education for American Democracy. Volume I. Establishing the Goals. New York: Harper and Brothers. 1947. pp. 57-58.

for all subject-matter courses. For science courses it is stated as a major objective. Problem-solving was mentioned as a specific objective of science teaching as early as 1920, when the report "Reorganization of Science in Secondary Schools"<sup>12</sup> suggested ways in which science instruction could contribute to the "Cardinal Principles of Secondary Education." In this report it was stated that useful methods of solving problems were specific values of the study of science.

The development of scientific attitudes was mentioned as one of the major objectives of science teaching in the Thirty-first Yearbook of the National Society for the Study of Education.<sup>13</sup> The Progressive Education Association lists the ability to think reflectively as one of the five broad areas of needs of adolescents.<sup>14</sup>

In "Science Education in American Schools," certain criteria were established for the formulation of objectives. The recommendations were made that objectives should be practicable for the classroom teacher. They also should be

---

<sup>12</sup> National Education Association, Reorganization of Science in Secondary Schools. U. S. Bureau of Education Bulletin, 1920, No. 26, Washington: Government Printing Office. pp. 12-15.

<sup>13</sup> Program for Teaching Science. Thirty-first Yearbook of the National Society for the Study of Education, Part I, p. 44. Bloomington, Illinois: Public School Publishing Company, 1932.

<sup>14</sup> Progressive Education Association. op. cit., p. 46.

psychologically sound, possible of attainment, universal in a democratic society and should indicate the relationship of classroom activity to the desired changes in behavior. On the basis of these criteria the committee suggested eight categories of objectives; one of these was problem-solving skills.<sup>15</sup>

That problem-solving skills are still one of the major objectives of the teaching of science is attested to by the fact that the Committee on Research in Secondary School Science of the National Association for Research in Science Teaching has set as one of its major tasks the identification of some of the important problems dealing with the teaching of problem-solving.<sup>16</sup>

Not only is the ability to solve problems a major objective of the secondary and elementary schools; but as shown by the following examples, it is also stated as a major objective of science teaching at the college level. The Harvard report<sup>17</sup> recommended that a part of the general education program in colleges be the teaching of an understanding of the means by which science has progressed.

---

<sup>15</sup> Science Education in American Schools. Forty-sixth Yearbook of the National Society for the Study of Education, Part I, pp. 19-40. Chicago: University of Chicago Press. 1947.

<sup>16</sup> Committee on Research in Secondary-School Science, "Problems related to the teaching of problem-solving that need to be investigated." Science Education, 34: 180-184, April, 1950.

<sup>17</sup> Harvard University, op. cit., pp. 220-230.

Gray<sup>18</sup> in 1931 listed "facility in application of the scientific method" as one of the objectives in the teaching of biology at the University of Chicago. In 1937, Greulack<sup>19</sup> in a committee report gave as one of the desired outcomes of biology teaching the development of scientific methods of thinking. To impart knowledge of the scientific method and encourage its use in thinking were listed as major objectives for the biology course at the University of Minnesota.<sup>20</sup>

Although the ability to think scientifically has been stated as a major objective of science by almost all educators there are still many unsolved problems in regard to this objective. In fact, as one considers the list of problems presented by the Committee on Research in Secondary-School Science<sup>21</sup> one wonders if anything at all is known about the teaching of the scientific method. The major problem areas considered by the committee were:

1. What is the nature of problem-solving in science?

---

<sup>18</sup> William S. Gray, editor, Recent Trends in American College Education. Chicago: University of Chicago Press. 1931. pp. 61-67.

<sup>19</sup> Muskingum College, A College Looks at its Program. Columbus: The Spahr and Glen Company. 1937. pp. 139-146.

<sup>20</sup> Ivor Spafford, editor, Building a Curriculum for General Education. Minneapolis: The University of Minnesota Press. 1943. pp. 245-261.

<sup>21</sup> Committee on Research in Secondary-School Science, op. cit., pp. 180-184.

2. How should problem-solving be taught?
3. How should ability in problem-solving be measured?

Approximately 150 problems were suggested by 53 of the members of the National Association for Research in Science Teaching who replied to a questionnaire concerning problems needing solving in the above areas. Some of the questions concerning the nature of problem-solving objectives in science teaching-learning situations which need to be answered and which are related directly or indirectly to the present investigation are:

- A. What are the specific skills and abilities necessary for successful problem-solving?
  1. Is problem-solving one ability or a composite of many different abilities?
  2. What are the fundamental components of the problem-solving ability?
  3. What is the relationship of problem-solving ability to general intelligence?
  4. Does the development of ability to solve problems depend chiefly upon the subject matter material or upon the manner in which it is presented?
- B. What is the relationship of individual differences in the following factors to the teaching of problem-solving?
  1. Ability to reason.
  2. Ability to read.
- C. What techniques can be used to measure a person's problem-solving ability?

1. Can the several kinds of problem-solving ability be expressed in any common measure?
2. Can the several components of problem-solving ability be appraised individually?
3. How can the validity of techniques for measuring problem-solving ability be established? Reliability?<sup>22</sup>

Almost all of the questions presented above are based on the assumption that there will be improvement in the ability to think scientifically if the teaching is directed toward that objective. But is this true? Some educators believe that the ability is an inherent one and that it does not yield to educative efforts. This point of view will be discussed more fully in Chapter II. Answers to most of the questions concerning methods of teaching scientific thinking, and the nature of scientific thinking depend upon a valid instrument to measure the ability to think scientifically. Although some tests have been devised to test certain abilities involved in scientific thinking, there are few if any tests now available which attempt to measure all of the inductive aspects of scientific thinking; nor are there any tests especially designed to measure these aspects of thinking for a course in first year college biology.

The present study is an outgrowth of an interest in writing laboratory studies for the laboratory guide used in

---

<sup>22</sup> Loc. cit.

Biological Science at Michigan State College which purports, among other things, to teach the student to think scientifically. Early in the evaluation of the effectiveness of the laboratory studies it became evident that until some measuring device for the ability to think scientifically was available no evaluation of the methods used in this laboratory guide was possible.

### THE PROBLEM

Statement of the problem. The purpose of this study was to devise a valid test to measure some of the inductive aspects of the ability to think scientifically.

The construction of test items required the identification of skills, and steps involved in scientific thinking, and the definition of behaviors which would give evidence of the ability to perform these skills. The validation of the test required the investigation of the relationship of whatever was measured by the test to (a) intelligence, (b) reading ability, (c) knowledge of biology, and (d) other measures of the ability to think scientifically, as evidenced by laboratory situations. In addition, it would require investigation to determine whether there was an increase in proficiency on the test after the completion of a course in biology which had as one of its major objectives the teaching of the ability to think scientifically.

Delimitation of the problem. The problem was limited to the construction of a test to measure the critical aspects of the inductive phases of scientific thinking. In this study the aspects of scientific thinking which were not creative activities, such as the sensing of a problem and the actual formulation of hypotheses, have been considered the critical aspects of thinking. A more detailed definition of these critical aspects of thinking and the reasons for limiting the test to the critical aspects will be discussed in Chapter IV. The reason for also limiting the test to the inductive phases was the fact that these phases of thinking were emphasized in the writing of laboratory studies for the course in Biological Science at Michigan State College. The items of the test were chosen from biological areas because the test was specifically devised for a course in first year biological science at the college level. No attempt has been made in this study to devise items to test the ability to apply principles of biology to new situations, nor has any attempt been made to construct items to test the attitudes which are assumed to attend the ability to think scientifically, namely, the scientific attitudes.

Basic assumptions of this study. The following are the major assumptions which underly this research.

1. Individuals differ in their ability to think scientifically.
2. These differences can be measured by direct



observation of the behavior of the individuals, and by indirect methods such as paper and pencil tests.

3. There are a number of skills involved in scientific thinking.

4. The behaviors which attend these skills can be described with sufficient objectivity to permit the devising of valid test items.

5. A sampling of an individual's reactions will give a measure of his reactions to a much larger range of situations.

6. The investigation of the ability to think scientifically is an important area of educational research.

Importance of the study. If the ability to think scientifically is an innate ability or if it is in reality general intelligence, as some educators believe,<sup>23, 24</sup> it is useless to attempt to attain it through the teaching of science. If, on the other hand, the ability is not innate or identical with general intelligence, as most educators believe, it should be teachable and it should be possible to determine which methods of teaching are most effective.

---

<sup>23</sup> Marion L. Billings, "Problem-solving in different fields of endeavor." American Journal of Psychology, 46:259-272, April, 1934.

<sup>24</sup> Ben D. Wood and F. S. Beers, "Knowledge versus thinking." Teachers College Record, 37:487-499, March, 1936.

In order to determine which of the above contrary opinions is correct, a test for the ability to think scientifically should be available.

#### ORGANIZATION OF THE REMAINDER OF THE THESIS

In Chapter II is presented a review of the research literature related to the problem. The first area of research reported is concerned with the identification of the steps involved in scientific thinking. The second portion of the review of literature is devoted to a discussion of tests which have been devised to measure various aspects of scientific or critical thinking. This discussion is followed by a review of research on the relationship of various aspects of critical thinking to such factors as intelligence, reading ability and knowledge of facts.

Chapter III is a discussion of the procedures involved in the development of a test designed to measure the ability to think scientifically.

Chapter IV is concerned with the steps involved in the development of the test items. The objectives, their definition in terms of desired behaviors, and illustrations of test items are included in this chapter.

Chapter V is concerned with the statistical analysis of the test and the test items. Item analysis data on the items of the preliminary tests and the statistical treatment of the preliminary and final forms of the test are presented.

Methods used to validate the test are presented in Chapter VI.

Chapter VII brings together the findings of this study with the conclusions to be drawn from them. This is followed by a discussion of the problems which the study has suggested and by the educational implications of the research.

## CHAPTER II

### REVIEW OF RESEARCH RELATED TO THE PROBLEM

In order to devise a test to measure the ability to think scientifically, it was necessary to determine the steps and skills involved in the use of the scientific method. Literature on this aspect of the problem is presented. This is followed by a review of tests which have been devised to measure various phases of scientific thinking. Previous work on the relation of the ability to think scientifically to various other characteristics such as intelligence, reading ability, and factual information is presented. A few studies on educability in ability to think scientifically are discussed.

### STEPS AND SKILLS OF SCIENTIFIC THINKING

Although much of a philosophic nature has been written on scientific method and individual scientists have described their methods of solving such problems, a review of these works has not been attempted here. Instead, the emphasis was placed on research aimed at determining the nature of this method. One exception was made in the case of Dewey, since he has been quoted frequently as an authority on problem-solving. The steps of problem-solving as conceived by Dewey<sup>1</sup> are:

---

<sup>1</sup> John Dewey, How We Think. Boston: D. C. Heath and Company. 1909. p. 72.

1. A felt difficulty.
2. Its location and definition.
3. Suggestion of possible solution.
4. Development by reasoning of the bearings of the suggestion.
5. Further observation and experimentation leading to its acceptance or rejection.

Until fairly recently little research had been done to determine the nature of the scientific method, although much has been written in the past 30 years on the desirability of teaching this method. Keeslar<sup>2</sup> surmised that the reluctance on the part of educators to investigate the steps of the method was due, (1) to the fact that problem-solving depends to some extent on the nature of the problem and, (2) to the tendency among researchers and writers to confuse the elements of the scientific method with scientific attitudes.

One of the earliest analyses of the elements of the scientific method was made by Downing<sup>3</sup> in 1928. For his steps in scientific thinking he drew upon illustrations from the history of science. In his list he included elements and safeguards of the scientific method. His safeguards were, in some instances, skills involved such as; inferences must

---

<sup>2</sup> Oron Keeslar, "A survey of research studies dealing with the elements of scientific method as objectives of investigation in science." Science Education, 29: 212-216, October, 1945.

<sup>3</sup> Elliot R. Downing, "The elements and safeguards of scientific thinking." Scientific Monthly, 26:231-243, March, 1928.

be tested experimentally and, in other cases, they were attitudes such as; judgment must be unprejudiced. It was Keeslar's<sup>4</sup> opinion that this failure to distinguish attitudes from elements has led to confusion of later workers and may have prevented a clear-cut definition of scientific method.

Tyler<sup>5</sup> discussed phases of scientific thinking in relation to the construction of tests to measure this ability. Davis,<sup>6</sup> LeSourd,<sup>7</sup> Downing,<sup>8</sup> and Beauchamp<sup>9</sup> described classroom techniques for the teaching of phases of scientific thinking. Curtis<sup>10</sup> analyzed the foregoing discussions and also incidents in the history of science. On the basis of these analyses he presented the following characteristics of scientific method as distinct from scientific attitudes.

---

<sup>4</sup> Keeslar, op. cit., p. 212.

<sup>5</sup> Ralph W. Tyler, Constructing Achievement Tests. Columbus, Ohio: Ohio State University. 1934. pp. 24-30.

<sup>6</sup> Ira C. Davis, "Is this the scientific method?" School Science and Mathematics, 34: 83-86, January, 1934.

<sup>7</sup> Homer W. LeSourd, "Teaching scientific method." School Science and Mathematics, 34: 234-235, March, 1934.

<sup>8</sup> Elliot R. Downing, "Teaching scientific method." School Science and Mathematics, 34: 400-405, April, 1934.

<sup>9</sup> Wilber L. Beauchamp, "Teaching scientific method." School Science and Mathematics, 34: 508-510, May, 1934.

<sup>10</sup> Francis D. Curtis, "Teaching scientific methods." School Science and Mathematics, 34: 816-819, November, 1934.

1. Locating problems.
2. Making hypotheses, or generalizations from given facts or observations.
3. Recognizing errors and defects in conditions or experiments described.
4. Evaluating data or procedures.
5. Evaluating conclusions in the light of facts or observations upon which they are based.
6. Planning and making new observations to find out whether certain conclusions are sound.
7. Making inferences from facts and observations.
8. Inventing check experiments.
9. Using controls.
10. Isolating the experimental factors.

In 1937, Crowell<sup>11</sup> prepared a list of 29 attitudes and 25 skills involved in scientific thinking. This list was derived from books and articles on philosophy, logic, science education, and science measurement. This list was presented to 64 science educators for evaluation. The skills rated as important by 80 percent of the judges are listed below in the order of their importance.<sup>12</sup>

1. Skill in observing accurately.
2. Skill in recording observations accurately and orderly.
3. Skill in forming independent judgments based on facts.
4. Skill in distinguishing between a fact and a theory.
5. Skill in picking out pertinent elements from a complex situation.
6. Skill in recognizing errors and defects in conditions and processes.
7. Evaluating conclusions in the light of facts or observations on which they are based.
8. Isolating the experimental factor.

---

<sup>11</sup> Victor L. Crowell, Jr. "The scientific method." School Science and Mathematics, 37:525-531, May, 1937.

<sup>12</sup> Loc. cit.

9. Forming sound judgments concerning adequacy of data.
10. Synthesizing or putting together separate facts to form a conclusion.
11. Gathering data systematically.
12. Planning an experiment to determine whether or not a proposed hypothesis is true.
13. Evaluating data and procedures.
14. Recognizing omissions or deficiencies in set ups.
15. Profiting from worthwhile criticism (an attitude?).
16. Forming a reasonable generalization.
17. Arranging and classifying data in sequence and making conclusions obvious.
18. Applying general principles to a new situation.
19. Recalling selectively items essential to a problem.
20. Locating problems.
21. Disregarding irrelevant facts.
22. Directing imagination into new and worthwhile channels.
23. Using the scientific instruments common in the laboratory.

Although 23 skills were rated as important by 80 per-cent of the respondees, no attempt was made to organize these skills into a plan for over-all problem-solving techniques.

Until 1945, when Keeslar<sup>13</sup> reported his study on the elements of scientific method, no adequately validated list of these elements was available. His original list of elements of scientific method was prepared on the basis of a survey of 43 books and articles on the scientific method. This list was then presented for validation to 22 research scientists at the University of Michigan. Elements considered to be of minor importance by the judges were eliminated from the list. The 42 remaining items were considered and

---

<sup>13</sup> Keeslar, op. cit., pp. 212-216.



combined, and were reorganized to form a final list of 10 major and 17 minor elements set forth in the order in which they might logically be expected to occur in the solution of a problem. This list was then checked by three specialists in the teaching of science.

The following is Keeslar's<sup>14</sup> list of major and minor elements of scientific thinking:

- I. Sensing a problem and deciding to try to find the answer to it. (italics in the original)
- II. Defining the problem. (italics in the original)  
     Stating the problem in words.  
     Analyzing the problem into its essential factors.
- III. Studying the situation for all facts and clues bearing upon the problem. (italics in the original)  
     Drawing upon past experience, both personal and those reported in literature, for possible explanations or generalizations to account for the phenomena observed.
- IV. Making the best tentative explanations or hypotheses as to the possible solution of the problem. (italics in the original)  
     Recognizing the assumptions which must be made if one goes beyond the known facts in formulating a hypothesis.
- V. Selecting the most likely hypothesis. (italics in the original)
- VI. Inventing and carefully planning one or more experiments to test the hypothesis, isolating the experimental factor wherever possible by using a control. (italics in the original)  
     Deciding upon the kinds of evidence which should be collected.

---

<sup>14</sup> Loc. cit.

Choosing reliable methods of collecting the evidence.

Refining measuring instrument to the degree warranted by the nature of the problem.

Practicing to gain skill in manipulation in order to secure accurate results.

VII. Testing the hypothesis by carrying out the experiment with great care and accuracy. (*italics in the original*)

Preventing, as far as possible, all uncontrolled variations in the conditions which might affect the results.

Making quantitative measurement of experimental results and estimating the probable error of such measurements.

Recording the results, adhering strictly to standard definitions and usage of scientific terms.

Organizing the pertinent data so that they may be studied and summarized.

VIII. Running check experiments involving the same experimental factor to verify the results secured in the original experiment. (*italics in the original*)

Studying the condition of the experiment in order to detect any omissions, defects, or errors, particularly those errors which might have been introduced in the experimental results by coincidence or chance.

Recognizing and, if possible, checking further the validity of the assumptions involved in setting up the experiment.

IX. Drawing a conclusion. (*italics in the original*)

Arriving at a solution to the problem based on an honest, unbiased appraisal of the data.

Suspending judgment when results are not conclusive.

Calling attention in the conclusion to those basic assumptions which it has been necessary to maintain throughout the procedure.

X. Making inferences based on this conclusion when facing new situations in which the same factors are operating.<sup>14</sup> (*italics in the original*)

---

<sup>14</sup> Keeslar, loc. cit.

Keeslar<sup>15</sup> concluded that the elements of the scientific method are definite, are distinct from attitudes, and are known and used by scientists. There was a high degree of agreement among the research scientists concerning the nature of these elements, thereby indicating that the scientific method has developed beyond the introspection stage and that teaching and testing can be based upon these skills. The 46th Yearbook<sup>16</sup> presented a somewhat more comprehensive list of skills than Keeslar's. Apparently it was based on Keeslar's list plus additions from various other sources.

The foregoing discussion has presented a brief survey of the research which has led to a definition of scientific method. It is interesting to note that the steps conceived by Dewey<sup>17</sup> in 1909, were basically the same as those derived from research in this area.

#### THE MEASUREMENT OF PROBLEM-SOLVING ABILITIES

In the last three decades a number of tests have been devised to measure various phases of scientific thinking. Some of these tests purported to measure numerous behaviors

---

<sup>15</sup> Loc. cit.

<sup>16</sup> Science Education in American Schools. Forty-sixth Yearbook of the National Society for the Study of Education, Part I, pp. 145-147. Chicago: The University of Chicago Press, 1947.

<sup>17</sup> Dewey, op. cit., p. 72.

while others were designed to measure very specific behaviors; such as, the ability to interpret data, or the ability to plan experiments. The following discussion presents the historical sequence of the tests which have been devised and the techniques which have been used to appraise the abilities involved.

As Glaser<sup>18</sup> has pointed out, several of the abilities included under the concept of the ability to think critically are, to some extent, measured by intelligence tests. Although such tests may be related in general to tests of scientific thinking no attempt will be made in this review to include tests or parts of tests which purport to measure general intelligence or any of its aspects.

Tests and scales have been devised to measure both the skills involved in problem-solving and the attitudes which attend these abilities. Some purported to measure both skills and attitudes while others, which were called attitude tests, contained some of the skills involved in scientific thinking. This review of tests will be limited to those which seem to measure skills involved in scientific thinking, and will not include tests and scales that measure attitudes only.

One of the earliest tests devised to measure the ability

---

<sup>18</sup> Edward M. Glaser, An Experiment in the Development of Critical Thinking. Contributions to Education, No. 843. New York: Bureau of Publications, Teachers College, Columbia University. 1941. p. 73.

to think scientifically was published in 1918 by Herring.<sup>19</sup> On the basis of an analysis of the work of such men as Francis Bacon, John Stuart Mill, and Karl Pearson, Herring selected eleven processes which he believed could be evaluated by a test. Herring stated that all of his eleven processes together did not constitute the whole of the scientific method, but he did believe that they all fell within the concept. His eleven processes, expressed in terms of the abilities involved, were (1) value, (2) feasibility, (3) definition, (4) clarity, (5) statistics, (6) relevancy, (7) recording, (8) comparison, (9) classification, (10) arrangement, and (11) sufficiency.

The test was devised for elementary and high school classes in geography. It contained thirty-three items of the multiple choice type. A direction was given which was followed by twelve choices. A thirteenth choice was available to indicate that none of the twelve choices were satisfactory.

The test was validated by being submitted to six judges. The judges indicated the answers they considered to be the correct ones and judged the fitness of the items as measures of the abilities which they were supposed to

---

<sup>19</sup> John P. Herring, "Measurement of some abilities in scientific thinking." Journal of Educational Psychology, 9:535-558, December, 1918.

measure. Estimates of the reliability of the test were not given. An interesting point about the test was that the processes described were expressed in terms of the abilities to be measured.

In 1924, Curtis<sup>20</sup> devised a test to measure the values derived from extensive reading in general science. It was designated as an attitude test and purported to measure, (1) a conviction of the universality of cause and effect relations, (2) the habit of delayed response, (3) the habit of weighing evidence with respect to pertinence, soundness, and adequacy, and (4) respect for another's point of view. The test was comprised of 34 items; some short answer items, and some multiple choice items. No reliabilities were given for the test.

Watson,<sup>21</sup> in 1925, published a test of fair-mindedness which purported to measure prejudice. In reality, this test probably measured much more than prejudice. The test was made up of six different types of sub-tests, some of which seemed to be measures of prejudice while others appeared to be measures of ability to think critically. A description of his

---

<sup>20</sup> Francis D. Curtis, Some Values Derived from an Extensive Reading of General Science. Contributions to Education, No. 163. New York: Bureau of Publications, Teachers College, Columbia University. 1924. pp. 57-67.

<sup>21</sup> Goodwin B. Watson, The Measurement of Fairmindedness. Contributions to Education, No. 176. New York: Bureau of Publications, Teachers College, Columbia University. 1925. pp. 9-35.

six sub-tests follows:

1. Form A was a list of 51 words. Instructions were given to cross out annoying or distasteful words.
2. Form B presented 53 statements about religious or economic matters upon which authorities differ. Instructions were given to mark each statement as true, probably true, uncertain or doubtful, probably false, or false. This type of key has probably been used more frequently in tests devised to measure abilities involved in critical thinking than any other single type of answer key. Watson's<sup>22</sup> test seems to be the first one in which it was used.
3. Form C, entitled the Inference Test, presented statements of fact followed by conclusions which might be drawn from the facts. Instructions were given to check only inferences which were certain and not to check those which were merely probable. One of the alternative answers was that no such conclusion could fairly be drawn. In each case one of the conclusions was a restatement of the data. In each case the only answers considered correct were the restatement of the data or the response that no conclusion could be drawn.
4. Form D was a moral judgments test. Fifteen instances of behavior were presented to be judged.
5. Form E was an arguments test based on the

---

<sup>22</sup> Watson, loc. cit.

assumption that a person will tend to feel that all arguments on the other side are weak. Twelve issues were presented followed by arguments.

6. Form F, the Generalization Test, contained unwarranted generalizations about groups as a whole. Subjects were asked to indicate whether the statement was true for all, most, many, few, or no individuals of the group. This test was scored on a negative basis, that is, a high score indicated that a person was not fairminded, a low score indicated that he was fairminded. The estimate of the reliability, determined by the split-half method was .96. The test was validated by:

1. Examination of the tests with reference to what they seemed to be measuring.
2. A study of the scores obtained by persons who were considered by their groups to be fairminded. This group actually had a lower average score than an unselected group (indicating fairmindedness).
3. A study of individuals who were supposed to be prejudiced by persons who knew them well.
4. A study of groups who would be suspected of certain lines of prejudice.
5. A correlation of test scores with other test scores. Results showed almost zero correlation both with reading test scores and with intelligence test scores.<sup>23</sup>

In the same year in which Watson described his test,

---

<sup>23</sup> Watson, loc. cit.



Daily<sup>24</sup> described a test to measure the ability of high school pupils to select essential data in solving problems. The test was not an objective test, but has been included here because it seems to be one of the first tests devised to measure a student's ability to recognize insufficiency of data and ability to select pertinent data. Eighteen short paragraphs containing data were presented. In some cases the data were insufficient; in other cases there were superfluous data. The student was asked to answer questions; the answers were, in reality, conclusions based on the data. Daily<sup>25</sup> reported the reliability of the test to be .73. The reliability was estimated by presentation of the same test seven weeks after the first administration of the test.

The Stanford Scientific Aptitude test was devised in 1927 by Zyve<sup>26</sup> to satisfy a need for more accurate guidance of incoming college students. It has been called an aptitude test because Zyve claimed that it tested inherent ability of the individual and not his achievement. The test included eleven elements of scientific aptitude; namely, (1)

---

<sup>24</sup> Benjamin W. Daily, The Ability of High School Pupils to Select Essential Data in Solving Problems. Contributions to Education, No. 190. New York: Bureau of Publications, Teachers College, Columbia University. 1925. pp. 59-60, 90-96.

<sup>25</sup> Loc. cit.

<sup>26</sup> D. L. Zyve, "A test of scientific aptitude." Journal of Educational Psychology, 18:525-546, November, 1927.

experimental bent, (2) clarity of definition, (3) suspended versus snap judgment, (4) ability to reason, (5) ability to detect inconsistencies, (6) ability to detect fallacies, (7) induction, deduction and generalization, (8) caution and thoroughness, (9) discrimination of values in selecting and arranging experimental data, (10) accuracy of interpretation, and (11) accuracy of observation.

The estimated reliability of the test was .93. The test was validated by having two judges rank students according to their aptitude for science. These rankings were compared with the rank of the students in their test performance. The coefficient of correlation between the scores on the Stanford Scientific Aptitude Test and the ratings of the judges was .74. The means of the test for science and engineering students and for a science faculty group were considerably higher than the means of a group of entering freshmen and non-science faculty.

Zyve's<sup>27</sup> test appears to be one of the first tests to make a successful attempt to measure scientific ability. Whether the test measures innate aptitudes which it purports to measure, or whether it measures an ability which can be learned does not seem to have been investigated despite the fact that the test has been rather widely used.

---

<sup>27</sup> Zyve, loc. cit.

Hoff<sup>28</sup> devised a scientific attitude test in 1930 which included the habit of weighing evidence as one of the attitudes measured. The test was validated by fifteen expert judges and by correlation with intelligence test scores and reading scores. These correlations were positive but low. The reliability given was .76, calculated by the split-half method.

A test of scientific thinking was published by Downing<sup>29</sup> in 1936, but had been used as early as 1931 by Strauss.<sup>30</sup> The test was designed to measure skill in the use of fifteen elements and safeguards involved in scientific thinking. The items were designed to test:

1. Accuracy of observation.
2. Ability to pick out pertinent elements from a complex situation.
3. Ability to synthesize.
4. Selective recall.
5. Fertility of hypotheses.
6. Ability to define a problem before trying to solve it.
7. Ability to hold in mind a complex of relations.
8. Problem-solving ability.
9. Judgment on adequacy of data.
10. Tendency to try to solve a problem scientifically rather than by trial and error.
11. Tendency to suspend judgment on moot questions.
12. Ability to apply a rule or law.

---

<sup>28</sup> Alfred G. Hoff, "A Test for Scientific Attitude." Unpublished Master's thesis, Department of Education, University of Iowa, 1930. pp. 1-42.

<sup>29</sup> Elliot R. Downing, "Some results of a test on scientific thinking." Science Education, 20:121-128, October, 1936.

<sup>30</sup> Sam Strauss, "Some results of the test of scientific thinking." Science Education, 16:89-93, December, 1931.

13. Tendency to test an hypothesis by collecting facts.
14. Awarments of the danger of reasoning by analogy.
15. Ability to arrange data in sequence to make the conclusion evident.<sup>31</sup>

As determined by the split-half method, the reliability of the test was .99 for a group of eighth through twelfth grade students. In general, each of the abilities tested was measured by a single question. Glaser<sup>32</sup> has criticized this test from the point of view of sound test construction and raises serious questions concerning its reliability and validity.

In 1933, Weller<sup>33</sup> constructed a test of 21 items which was designed to measure the effectiveness of teaching of scientific thinking in the elementary schools. Seven sets of items were used. The first item of each set attempted to measure observation, the second item asked the student to draw a conclusion from simple data, and the third item asked for a proof or possible verification of the conclusion drawn. She found the reliability of this portion of her test to be .54.

Noll,<sup>34</sup> in 1933, described a test of scientific

---

<sup>31</sup> Downing, op. cit., pp. 121-128.

<sup>32</sup> Glaser, op. cit., p. 76.

<sup>33</sup> Florence Weller, "Attitudes and skills in elementary science." Science Education, 17: 90-97, April, 1933.

<sup>34</sup> Victor H. Noll, The Habit of Scientific Thinking, A Handbook for Teachers. New York: Bureau of Publications, Teachers College, Columbia University. 1935. pp. 18-25.

thinking entitled, "What do You Think?" The test was constructed to satisfy a need in the schools for a test to evaluate the teaching of scientific thinking.

Six habits of thinking were selected as a basis for constructing the preliminary forms of the test. Each question was intended to express a situation which was familiar to most persons, and which afforded an opportunity for scientific thinking. The preliminary form of the test included 134 items, most of which were of the true-false type. Approximately 25 items were designed to measure each of the six habits of thinking, namely; accuracy of observation, intellectual honesty, openmindedness, suspended judgment, a conviction of universal operation of the law of cause and effect, and criticism.

The reliabilities of the two final forms of the test were determined in two ways. The method of split-halves corrected by the Spearman-Brown formula gave a reliability of .82 for Form I and a reliability of .92 for Form II. A correlation between the two forms of the test gave a reliability of .69. Noll<sup>35</sup> believed that the true reliability coefficient was probably somewhere between the highest and the lowest figures obtained.

The test was validated by correlation with I.Q.'s and by the determination of item validity. The correlation of

---

<sup>35</sup> Noll, loc. cit.

the test with I.Q.'s ranged from .30 to .41, indicating that native ability was not being tested to a large extent. Norms for grades eight through twelve were presented.

In 1936 Frutchey, Tyler and Hendricks<sup>36</sup> reported a test to measure the ability to interpret experimental data. This report is of interest, not because it presents the construction of a complete test, but because it reports an investigation of the validity of a particular type of item. In Test I an experiment was described and the student was asked to write a conclusion. Although this is, in some ways, a very satisfactory method of evaluating a student's ability to draw conclusions, it is difficult to grade; therefore, Test II was prepared. The same experiments were used and five conclusions were selected from the free responses of students in Test I. The students were instructed to select the best conclusion. This method did not give a valid measure of a student's ability to formulate conclusions since the correlation of the scores on Test I and Test II was only .38. This same test was rendered more valid when the student was asked to check the best conclusion and the one contradicted by the data. This was designated as Test III. It's correlation with Test I was .85. In the final form of the test which

---

<sup>36</sup> Fred P. Frutchey, Ralph W. Tyler and B. Clifford Hendricks, "Measuring the ability to interpret experimental data." Journal of Chemical Education, 13: 62-64, February, 1936.

proved to be the most valid, the same test items were used but students were instructed to mark each item according to the following key:

Mark with a 1 every statement which is a reasonable interpretation of the data.

Mark with a 2 every statement which might possibly be true but for which insufficient facts are given to justify the interpretation.

Mark with a 3 every statement which cannot be true because it is contradicted by the results obtained in the experiment.

Love,<sup>37</sup> in 1937, devised a test of scientific attitudes and scientific thinking. The test was in three parts and contained 24 items. Part I dealt with the criticizing and planning of experiments; Parts II and III tested the ability to recognize assumptions upon which conclusions were based.

Raths,<sup>38</sup> in 1938, described a test designed to evaluate thinking ability. The first portion of the test dealt with interpretation of data. The student was required to determine the probable truth or falsity of a series of statements concerning the data. The second portion of the test contained a description of a situation followed by three conclusions. The students were instructed to choose the best

---

<sup>37</sup> Kenneth G. Love, "Scientific Attitude - Thinking." Every Pupil Test. Columbus, Ohio: The State Department of Education. April. 1937.

<sup>38</sup> Louis E. Raths, "Evaluating the program of a school." Educational Research Bulletin, 17:57-84, March, 1938.

conclusion. The conclusions were followed by a series of reasons which could be used to explain why the conclusion was chosen. The students were instructed to indicate the reasons they had chosen a particular conclusion. The third portion of the test presented a situation and a conclusion based upon the situation. These were followed by a series of statements some of which were assumptions. The student was instructed to check the assumptions and to indicate those upon which the conclusion was based. He was then required to organize a proof for the conclusion using the assumptions and data. No reliabilities for the test were given.

The tests devised by the evaluation staff of the Eight-Year Study were published in 1938, and were described in detail by Smith and Tyler<sup>39</sup> in 1942. The Eight-Year Study<sup>40</sup> was planned to implement broad objectives of education in the secondary schools without regard to college entrance requirements. The experiment was confined to thirty selected secondary schools throughout the United States. Students from these schools were admitted to colleges on the basis of recommendation by the principal of the school and

---

<sup>39</sup> Eugene R. Smith, Ralph W. Tyler, and the Evaluation Staff, Appraising and Recording Student Progress. New York: Harper & Brothers. 1942. pp. 3-156.

<sup>40</sup> Wilford M. Aikin, The Story of the Eight-Year Study. New York: Harper and Brothers. 1942. pp. 12-24.



not on the basis of college entrance requirements or examinations. Extensive studies of objectives and means of evaluation of objectives were, among other things, a part of this project. The behaviors which were to be measured by the tests were defined by committees composed of the members of the evaluation staff of the Eight-Year Study<sup>41</sup> and representatives from each school interested in the objectives being measured. Two of the objectives related to the present study were, the ability to interpret data, and the ability to understand the nature of proof.

The earlier forms of the interpretation of data tests were intended primarily for use in the senior high school. Ten sets of data, presented in various forms including prose, graphs, tables, and charts were each followed by 15 statements. The students were instructed to evaluate each of these on the basis of the following key:

- (1) are sufficient to make the statement true.
- (2) are sufficient to indicate that the statement is probably true.
- (3) are not sufficient to indicate whether there is any degree of truth or falsity in the statement.
- (4) are sufficient to indicate that the statement is probably false.
- (5) are sufficient to make the statement false.<sup>42</sup>

---

<sup>41</sup> Smith and Tyler, op. cit., pp. 3-156.

<sup>42</sup> Ibid., p. 52.

In the early history of the development of tests to measure the ability to interpret data, tests were devised for specific subject matter fields. However, the evaluation staff believed that the behaviors involved in these tests were not essentially different so a single measuring instrument was constructed. In all, nine forms have been used; the last two, Interpretation of Data Test Form 2.51 and 2.52, have been prepared as alternate forms. Forms 2.71 and 2.72 were prepared for use in the junior high schools.

The answers to the test items were validated by the judgment of a group of experts in the field and by preliminary tryouts on groups of students.

The method of scoring these tests is of considerable interest. The tests were scored four separate times to give the following scores:

1. General accuracy score was the total number of answers which agreed with the answers of the jury of experts. This score was expressed as the percent of the maximum possible number of correct responses.

2. The "going beyond data" score was calculated by determining the number of times a student considered a statement to be true which the jury had considered only probably true, or probably true when the jury had considered it as insufficient data, etc.

3. The "caution" score indicated the extent to which a student marked statements keyed true as probably true; keyed probably true as insufficient data, etc.

4. The "crude error" score was obtained by determining the extent to which students marked items in contradiction to the data.<sup>43</sup>

---

<sup>43</sup> Ibid., pp. 54-55.

The tests on interpretation of data were validated; (a) by comparing the behaviors demanded of students in the test with the behaviors defined in the statement of objectives to be measured, (b) by selecting data which were of the type which students encounter in textbooks, and (c) by studying the distribution and means of scores made by students in various grades of school. The means increased with grade levels. Another method used in the validation of the tests was the comparison of test scores with essay responses on the same data.

The reliabilities of the various types of scores on Form 2.52 of the test computed by use of the Kuder-Richardson formula ranged from .81 to .95. The general accuracy score was the most reliable. The split-halves method of estimating reliability was used for Form 2.51. Reliabilities ranged from .86 to .92. Comparisons of the two forms yielded reliability coefficients of from .65 to .85.<sup>44</sup>

Another of the tests devised by the Evaluation Staff of the Eight-Year Study<sup>45</sup> was the "Nature of Proof." This test was devised to measure the ability of students to locate and appraise the basic assumptions upon which the proof of a statement depended. A paragraph containing data was followed by a conclusion. Following this were 14 statements, some of

---

<sup>44</sup> Ibid., pp. 65-76.

<sup>45</sup> Ibid., pp. 128-154.

which were assumptions underlying the argument. In the first part of the test, the student was asked to decide which statements were relevant to the conclusion and to mark them as either supporting or contradicting the conclusion. In the second part of the test, the student was asked to indicate which of the statements marked as supporting the conclusion he would challenge. In the third part of the test, the student was instructed to choose one of three stated conclusions. In the fourth part, the student was asked to select activities which might be useful in the solution of a problem related to the previous conclusions. In part five the student was directed to indicate which of these activities could be carried out in a school situation. Reliabilities of the various part scores on the test ranged from .20 to .82.

Two interesting types of items devised to measure critical thinking in a science course have been described by Hart.<sup>46</sup> A statement of a situation was presented. This was followed by a numbered series of observations. Five hypotheses were then presented and the student was instructed to list by numbers all of the observations which supported each of the hypotheses. He was also instructed to list by number all of the facts which weakened each hypotheses. Then

---

<sup>46</sup> E. H. Hart, "Measuring critical thinking in a science course." California Journal of Secondary Education, 14:334-338, October, 1939.

the most valid hypothesis was to be checked. The second type of item described was similar to the above but an hypothesis was chosen by the student before the data were presented. The data were used to support, weaken, or eliminate the hypothesis. No data on validity or reliability of tests composed of such items were given.

In 1940, Gans<sup>47</sup> described a test used in a study of critical reading comprehension. The test was devised to measure ability to recognize problems and to solve problems by critical selection and rejection of data. Paragraphs containing problems were presented. An item followed which had as its foils three problems. The student was asked to determine which problem had been presented in the paragraph. The problem item was followed by a series of paragraphs containing facts which were directly related, indirectly related, or unrelated to the problem. The student was instructed to mark each paragraph according as to whether it did or did not aid in the solution of the problem. These paragraphs were followed by a three-choice item asking for the major problem under consideration. This was followed by single statements of facts taken from the paragraphs previously presented. Again, the student was requested to indicate whether the fact

---

<sup>47</sup> Roma Gans, A Study of Critical Reading Comprehension in the Intermediate Grades. Contributions to Education, No. 811. New York: Bureau of Publications, Teachers College, Columbia University. 1940. pp. 59-89.

helped or did not help in the solution of a problem. In addition, he was asked to judge the truth or falsity of each of these statements.

The test was scored as five subtests. The reliabilities of the subtests ranged from .67 to .90. The total test reliability was not given.

Engelhart and Lewis,<sup>48</sup> in 1941, described a 23 item portion of a pretest for a physical science survey course at Chicago City Junior College. These 23 items were designed to measure scientific thinking. In an introductory paragraph the terms hypothesis and conclusion were defined. An experimental situation was described, a problem was stated, and the following key was presented:

"Below are given a series of hypotheses, each of which is followed by numbered items which represent data. After each item number on the answer sheet blacken space -

- A. if the item directly helps to prove the hypothesis true.
- B. if the item indirectly helps to prove the hypothesis true.
- C. if the item directly helps to prove the hypothesis false.
- D. if the item indirectly helps to prove the hypothesis false.
- E. if the item neither directly nor indirectly helps to prove the hypothesis true or false."<sup>49</sup>

---

<sup>48</sup> Max D. Engelhart and Hugh B. Lewis, "An attempt to measure scientific thinking." Educational and Psychological Measurement, 1:289-294, Third Quarter, 1941.

<sup>49</sup> Loc. cit.

Three hypotheses were presented, each hypothesis was followed by five statements of fact. These statements constituted the items, which were marked by the above key. These items constituted 15 items of the test. The student was directed to judge each hypothesis as to its truth or falsity. These judgments constituted three items of the test. Following these 18 items five conclusions were given. Each conclusion was to be judged either the best, the worst, or neither best nor worst.

The items of this test proved to be quite discriminating, the range of correlations of items with the total score on the 23 items being from .17 to .61. The reliability of the test was estimated to be .72 by means of the Kuder-Richardson formula.

The Watson-Glaser Tests of Critical Thinking were described by Glaser<sup>50</sup> in 1941. These tests were designed to appraise some of the abilities involved in critical thinking. They were, in effect, an extensive revision of Watson's tests of fair-mindedness. All of the tests were validated by 15 judges.

Test A, A Survey of Opinions, was devised primarily to show the extent of a person's consistency of opinion. The test-retest reliability was .88; the correlation between scores on Section I of the test and Section II of the test

---

<sup>50</sup> Edward M. Glaser, op. cit., pp. 87-92.

was .85.

Test B, the General Logical Reasoning Test, was designed to measure the ability to think in accord with the rules of logic. The test-retest coefficient of reliability was given as .82.

Test C, the Inference Test, was designed to measure ability to judge the probable truth or falsity and the relevance of inferences drawn from given facts. The persons taking the test were instructed to determine whether the conclusions drawn were true, probably true, false, probably false, or questionable. The test was validated by the fact that the test significantly distinguished between two groups of students judged by their teachers to be either superior or inferior in ability to think logically. Test-retest reliability was found to be .86.

Test D, the Generalization Test, was substantially the same as the one of the same name devised by Watson and discussed earlier in this review. The reliability of this test was reported as .88.

Test E, the Discrimination of Arguments, was also substantially the same as the arguments test of Watson's earlier edition. The reliability given for this test was .76.

Test F, the Evaluation of Arguments Test, was a new test in the series. Each test item consisted of a paragraph followed by three alternative conclusions, only one of which was logical on the basis of the data presented in the



paragraph. Following the conclusions six reasons were listed, one of which explained why the correct conclusion was the logical one. The testee was instructed to check the reason explaining his conclusion. The test-retest coefficient of reliability for this test was .83.<sup>51</sup>

Fleming,<sup>52</sup> in 1942, described a test used in his analysis of outcomes of a course in biological science. A portion of the test was devoted to the measurement of the ability to think scientifically. The items for the test had been chosen from examinations given previously in the course. This portion of the test was divided into four parts; Part A was designed to measure the recognition of steps in problem solving, Part B was an evaluation of statements with reference to a problem, Part C was designed to measure the ability to evaluate inferences, and Part D was the selection of data pertinent to the solution of a problem situation. The tests were described but no test items were included in Fleming's dissertation. The reliability of this portion of the test was not given.

A test designed to measure a student's ability to judge conclusions was constructed by Higgins<sup>53</sup> in 1942 to

---

<sup>51</sup> Glaser, loc. cit.

<sup>52</sup> Maurice C. Fleming, "An Analytical Study of Certain Outcomes of a Course for Orientation in Biological Science at College Level." Unpublished Doctor's thesis, Department of Education, New York University, 1942. Appendix.

<sup>53</sup> Conwell D. Higgins, "Educability of Adolescents in Inductive Ability." Unpublished Doctor's thesis, Department of Education, New York University, 1942. pp. 36-40, 133-137.

evaluate educability in inductive ability. Twelve experiments were described; each experiment was followed by a series of conclusions which constituted the items. There were a total of 97 items which had been selected from free responses of students. The testees were instructed to determine whether the conclusions were complete, incomplete, based on insufficient data, or false. The test was validated by agreement of four judges as to the correct answers to the items. The estimate of reliability, as determined by the split-half method, was .90.

Ter Keurst and Bugbee,<sup>54</sup> in 1943, published a test by which the authors claim "teachers or students can check themselves on the understanding of the methodology of science." The test consists of a series of four-choice items which purport to measure knowledge of skills, attitude, and terminology of scientific method. The test seemed to test knowledge but not behavior. However, it is of interest to note, that it apparently had a certain degree of validity. The test was administered to a group of students who had been named as the five best and the five worst students in science classes in respect to their ability to think scientifically. The critical ratio of the difference of the means of these two groups was 5.01. The test was also validated by opinion of experts.

---

<sup>54</sup> Arthur J. Ter Keurst and Robert E. Bugbee, "A test on scientific method." Journal of Educational Research, 36:489-501, March, 1943.

The estimate of the reliability by means of the split-half method was .82.

A very interesting test in two forms entitled, "Do You Think Straight?" was described by Johnson,<sup>55</sup> in 1943. The test was designed to measure the relation of reflective thinking to ability in debating and discussion. Because her test was an attempt to overcome some of the inadequacies of earlier tests her criticisms of existing tests are presented here:

These tests, though useful, appear to be inadequate for the diagnosis and measurement of the process (italics in the original) of reflective thinking. Each test is deficient on two or more of the following counts:

1. It breaks the process of reflective thinking into what may be superficially (italics in the original) distinct and uncoordinated units.
2. Even in measuring such units, the following factors or steps are not considered:
  - a. The formulation of a problem.
  - b. The analysis into major variables.
  - c. The determination of criteria and application of them to the evaluation of possible solutions.
  - d. The construction and comparison of hypotheses.
3. It deals with a great variety of problems - each item relating to a different problem, in most tests - whereas the need in actual life situations (and the need in discussion and other forms of public speaking) is to think through (italics in the original) a particular problem.
4. It emphasizes the logic of intentional (italics in the original) reasoning - the discrimination among formally valid and invalid conclusions and "reasons" for conclusions - rather than the logic of constructive (italics in the original) reasoning or scientific

---

<sup>55</sup> Alma Johnson, "An experimental study in the analysis and measurement of reflective thinking." Speech Monographs, 10: 83-96, Annual, 1943.

discovery. In fact, those tests which require the subject to check a conclusion and then to check reasons for his choice appear to be measuring little except expertness in "rationalizing."<sup>56</sup>

Johnson's tests were constructed on the assumptions that; (1) Dewey's steps were a correct description of the thought process, and (2) there were discoverable and observable obstacles to reflective thinking.

Forms A and B,<sup>57</sup> were each designed around a single problem. Section I of each test was an attempt to measure attitudes about the problem. In Section II, ten subsidiary problems were presented. The student was instructed to number these in order of their usefulness as starting points in the solution of the overall problem. Section III presented four groups of questions, each composed of three subordinate questions; the most important one to be checked. In Section IV, data were presented which might aid in the solution of the four major questions posed in Section III. These data were followed by statements which the student was instructed to mark as being true, probably true, insufficient data, probably false, or false. In Section V, ten syllogisms or pseudo-syllogisms were presented; followed by conclusions which the students were instructed to mark as sound or unsound. The students were instructed to rank the six solutions to the overall problem. This constituted Section VI of the

---

<sup>56</sup> Ibid., p. 85.

<sup>57</sup> Ibid., pp. 83-96.

test. Section VII required the matching of advantages and disadvantages, which were summaries of statements of information given throughout the test, with the three best solutions of Section VI. In the final section of the test, Section VII, the student was instructed to classify each of ten conclusions as critical, uncritical, hypercritical, or dogmatic.

Johnson stated that there was inherent validity in the test since it was patterned after Dewey's steps of thinking and since the syllogism test followed the rules of logic. In addition, however, the test was validated by 15 experts in the fields of logic and scientific method. She also found that scores of students judged superior in the abilities involved in reflective thinking were higher than those judged as average, and those judged average scored higher than those judged as inferior. She also cited an increase in scores in college grade levels as evidence for the validity of the test. These increases in scores with college grade levels were at the 5 percent level of significance or better.

The estimate of reliability, determined by correlating the scores made on the two forms of the test, was  $82 \pm .02$ . The scores on the attitude portion of the test were not included in the total test scores.

A portion of the test, which was used to appraise methods of teaching scientific method, was designed by

Thelen<sup>58</sup> to measure an understanding of experimental design. The purpose of an experiment was given; this was followed by conditions of the experiment and statements about the experimental material. The student was instructed to indicate which factor or factors were to be varied, which were to be fixed, which might be assumed to be negligible, which were irrelevant, and which factors the student did not understand. In all, there were 60 such items. No reliability was given nor was any evidence concerning the validity of the test presented.

In 1944, Raths<sup>59</sup> devised the "Ohio Thinking Checkup," a thinking test for students in the third, fourth, and fifth grades. Twelve problem situations were presented. Each problem was followed by eight statements which the students were instructed to mark as true, false, or questionable. Items were devised to reveal nine types of errors in thinking; namely,

1. Interpretation through personal judgment.
2. Evading of issue by name-calling or ridicule.
3. Leaning on authority.
4. Believing in superstition.
5. Generalizing from insufficient evidence.
6. Rationalizing or misinterpreting data.
7. Calling either-or statements true.

---

<sup>58</sup> Herbert A. Thelen, "An Appraisal of Two Methods for Teaching Scientific Method in General Chemistry." Unpublished Doctor's thesis, Department of Education, University of Chicago, 1944. pp. 365-369.

<sup>59</sup> Louis E. Raths, "A thinking test." Educational Research Bulletin, 23:72-75, March, 1944.

8. Calling if-then statements true.
9. Leaning on school loyalty.

The reported reliabilities of the tests as determined by the method of matched halves were .89 for the fourth grade, .91 for the fifth grade, and .93 for the sixth grade.

Grant and Meder,<sup>60</sup> in 1944, suggested a type of item to evaluate reasoning ability. A statement was presented followed by six reasons for agreeing with the statement and six reasons for disagreeing. The student was instructed to check valid reasons from either or both lists and then to decide whether he agreed or disagreed with the statement.

In 1944, reports of the high-school and the college chemistry tests for the armed forces were published. In each of these tests one section was devoted to items designed to measure abilities involved in scientific thinking. Ashford,<sup>61</sup> in reporting on the college test, listed six of these abilities which were to be measured. Items were devised to test the ability to (1) distinguish between observed phenomena and their theoretical explanation, (2) explain phenomena in terms of theory, (3) give the experimental evidence for a theory, (4) identify the assumptions

---

<sup>60</sup> Charlotte L. Grant and Elsa M. Meder, "Some evaluation instruments for biology students." Science Education, 28:106-110, March, 1944.

<sup>61</sup> Theodore A. Ashford, "The college chemistry test in the Armed Forces Institute." Journal of Chemical Education, 21:386-392, August, 1944.

necessary for a given conclusion, (5) identify the factor that must be controlled in an experiment, and (6) identify statements which are true merely by definition. The test was prepared in two forms; one for the armed forces, one for civilian use.

Hered and Thelen<sup>62</sup> devised a similar test for use at the high-school level. Single items were devised to measure each of the abilities which they had considered to be important in scientific thinking. The reliability coefficients of the tests were not given; however, Hered and Thelen reported that the reliability of the high-school test was satisfactory.

The ability of ninth grade students to make conclusions was investigated by Teichman.<sup>63</sup> For this investigation he designed three tests. Test A, which was not objective, presented 16 paragraphs from which the students were to draw conclusions. In Test B, 29 experiments were described; each was followed by four conclusions. The students were instructed to choose the best one. In Test C, 15 problems were presented, followed by data. A conclusion, which was faulty, was stated. These 15 faulty conclusions constituted the

---

<sup>62</sup> William Hered and Herbert A. Thelen, "The high-school chemistry test of the Armed Forces Institute." Journal of Chemical Education, 21:507-515, October, 1944.

<sup>63</sup> Louis Teichman, "The ability of science students to make conclusions." Science Education, 28:268-279, December, 1944.



items of Test C. Students were instructed to evaluate the faulty conclusions according to the following key:

- (a) It does not answer the problem or question.
- (b) It does not agree with the facts of the experiment.
- (c) There are not enough facts to make the conclusion valid (correct).
- (d) The facts have not been obtained by proper control (comparison) in the experiment.

The test was validated by unanimous agreement of three prominent educators in the field of science, by item analysis, and by intercorrelations of the three tests. The reliabilities were estimated by the split-half method. The reliability of Test A was .88, of Test B was .88, and of Test C was .68. The total test reliability was given as .91.

Alpern,<sup>64</sup> in 1946, devised a test for high-school students to measure the ability to suggest procedures to test hypotheses. From the responses of this non-objective test he constructed an objective test to measure the ability to select methods of testing hypotheses. Each of the test items consisted of (1) a situation, (2) a statement of the problem, (3) an hypothesis offered as an explanation, and (4) four suggested procedures. These last constituted the foils of each item; the student was instructed to choose the best experiment to test the hypothesis given. The

---

<sup>64</sup> Morris L. Alpern, "The ability to test hypotheses." Science Education, 30:220-229, October, 1946.

preliminary forms of this test were revised on the basis of criticism of experts and on the basis of item analysis.

Twenty items constituted the test which had an estimated reliability coefficient of .75. The test was validated by the judgment of 41 educators in science, by item analysis, by a consideration of the range of difficulty of the items, and by the fact that average scores increased through successive grades, from ninth through twelfth.

A test to measure certain aspects of scientific thinking in the area of college physics was devised by Dunning.<sup>65</sup> The test was constructed to measure ability to interpret data and ability to apply principles. The method of evaluation used by Dunning to test the ability to interpret data was substantially that reported by Smith and Tyler.<sup>66</sup> Dunning's unique contribution to the measurement of this objective was his use of four methods of scoring the papers in order to determine the effects of variously weighted scorings on the reliability. He found the method of giving a single point for the keyed answer gave the highest estimate of reliability by the split-half method. The reliability was given as .83. In addition, he found that this method also

---

<sup>65</sup> Gordon M. Dunning, "The Construction and Validation of A Test to Measure Certain Aspects of Scientific Thinking in the Area of First Year College Physics." Unpublished Doctor's thesis, Department of Education, Syracuse University, 1948.

<sup>66</sup> Smith and Tyler, op. cit., pp. 15-28.

gave the highest validity coefficient when he correlated scores on the test with teacher ratings of the students. The validity coefficient obtained was .56. A second method of validation of the test was the correlation of scores made on the objective test with scores on the same material on an essay test. This correlation was .66.

Ullsvik<sup>67</sup> constructed a test which was designed to measure critical judgment in geometry classes. The test, however, was on non-geometric subjects. The test was in three parts: Part I was called "Judging of Conclusions" and instructed the students to mark the conclusions given as acceptable, not acceptable, or insufficient evidence, Part II was an evaluation of definitions, Part III presented a paragraph followed by 15 statements. The student was instructed to select the two statements which were the most crucial in leading one to accept the conclusion, and the two which were the most crucial in leading one to reject the conclusion. The reliability of the test was not given.

In 1949, Read<sup>68</sup> published a description of a non-verbal test of the ability to use the scientific method. An

---

<sup>67</sup> Bjarne R. Ullsvik, "An attempt to measure critical judgment." School Science and Mathematics, 49:445-452, June, 1949.

<sup>68</sup> John G. Read, "A non-verbal test of the ability to use the scientific method as a pattern for thinking." Science Education, 33:361-366, December, 1949.

analysis of Keeslar's<sup>69</sup> major elements of scientific method led Read<sup>70</sup> to the inference that many of these steps involved discriminatory choices. The inventing and planning of experiments could only be measured by physical methods but the other elements he claimed all involve discriminatory choices. These were summarized as follows:

1. Observation is only valuable when it is discriminating.
2. The defining of a problem means a choice among possible problems.
3. Classification of data is discrimination between items.
4. Setting up hypotheses is the choosing of one or more possible explanations of the data.
5. Selecting the most likely hypothesis is critical discrimination.
6. Drawing conclusions is selecting and fitting of data, again critical discrimination.
7. Validation of the conclusion is again a matter of discrimination and choice.

On the basis of his contention that scientific thinking is primarily the making of discriminatory choices, he devised a picture-test to appraise the ability to make these choices. He described his test as follows:

The picture-test is a series of sub-tests, related in that they are all aspects of the environment, and that they all pose problems which can be solved through

---

<sup>69</sup> Keeslar, op. cit., pp. 212-216.

<sup>70</sup> Read, op. cit., pp. 361-366.

the association of two sets of pictures. There are seven categories; each edlinedated by four pictures, each of which represents a particular sub-division of the category. (Three more categories of a biological nature have been added). The categories have to do with electricity, with air pressure, with one phase of chemistry, with mechanics; they are samples of common environmental science.

The four pictures are mounted on a card, ..... the card is placed in a box. Under each of the four pictures is a small bin. From six to eighteen separate loose pictures may be picked up by the testee, closely examined, sorted, compared, and finally dropped into one of the bins. The only directions are to "place each picture in the bin where it fits best."

High scores are obtained by those who discover what the four pictures on the card represent. As each card is on a single topic, the task is to discover the more or less fine shades of *dis*-similarity (italics in the original) among the four pictures. The loose pictures serve as clues, and as they can be moved around without penalty, once the pattern exhibited by the four pictures on the card is discovered, the way is open for careful comparison and critical discrimination.<sup>71</sup>

Read originally used 133 pictures which he presented to eleven science specialists for sorting. Of these 133, seventy were placed by all of the judges in the same bins. Item-analysis showed that 27 of these were non-discriminatory; the remaining 43 pictures made up the items of the test. The test was designed for grades seven through twelve. By means of the Kuder-Richardson formula, Read found the reliability of the test to be .78. The test was validated by administering it to 18 members of the group who won high honors in a state science contest. The scores made by these students was significantly higher than scores

---

<sup>71</sup> Ibid., pp. 362-363

made by students who had had no science.

Bingham<sup>72</sup> devised a series of tests for general science, biology, chemistry, and physics which were used primarily as teaching devices. The instructor performed an experiment and then a twelve-item test was given. Item 1 was concerned with the results of the experiment. Item 2 described experiments; the student was directed to select the one actually performed. Item 3 presented five hypotheses to account for what happened; the student was instructed to choose the best one. In items 4-8 additional facts were given and the student was directed to choose the fact which showed the untenable hypotheses presented in Item 3 to be unsound. The choice, "none of these," could be used for the hypothesis which was sound. Item 9 tested an understanding of the assumptions underlying the conclusion drawn; Item 10 was concerned with new problems arising out of the experiment, while Item 11 presented assumptions underlying the application of the conclusion to new situations. Item 12 tested the ability to apply the conclusion to new situations. No data on the reliability or validity of the test were presented.

Edwards,<sup>73</sup> in 1950, reported on two tests, Test A and

---

<sup>72</sup> Eldred N. Bingham, "A direct approach to the teaching of the scientific method." Science Education, 33:241-249, April, 1949.

<sup>73</sup> Thomas B. Edwards, "Measurement of Some Aspects of Critical Thinking." Journal of Experimental Education, 18:263-279, March, 1950.

Test C, which he devised to measure certain aspects of critical thinking. Test A was devised to measure induction. Four principles were stated; each principle was followed by five facts. The pupil was instructed to choose the fact which supported the principle. The estimate of reliability of the test was .88 as determined by the method of split-halves, .80 as measured by a correlation of the two forms of the test. Edwards claimed that the validity was built into the test by using an accepted theory of critical thinking and by using facts familiar to students. Additional evidence for validity was found in an increase in scores from grades ten through grade fourteen (college sophomore) and in a correlation of only .17 with intelligence.

Test C was called a Judgment Test. Four opinions were stated; these were labeled A, B, C, and D. One opinion was sound, one fairly adequate, one irrelevant, and one totally incorrect. The opinions were then presented in pairs, AB, AC, etc., giving six items for each set of four opinions. The student was instructed to choose the better of each pair. This test was prepared in two forms. Reliability coefficients ranged from .49 to .75 when determined by the split-half method. The correlation between the two forms was .32. The methods of validation were the same as for Test A. The correlation of Test C with intelligence was .15.

Tests A and C were two tests of a battery of tests

devised by Edwards<sup>74</sup> who originally set out to measure seven aspects of critical thinking. Seven tests were devised. Test I aimed to test the ability to judge the reliability of sources of information. A series of statements concerning measurements were presented. The student was instructed to underline the letter R if he felt that the accuracy mentioned was possible by means of the device used, but to underline the letter N if the device could not measure as accurately as was indicated in the statement. Edwards states that this test showed some promise, but that it was not developed beyond the preliminary stages because the reliability was low.

Test II was a test of relevance. Each question consisted of two statements. The student was instructed to underline the letter R if the two statements were related, to underline the letter N if they were not related. This test was not revised after the first tryout because of the difficulty of obtaining facts which the test constructor was sure all of the students would know. Test III was the induction test discussed as Test A above. Test IV was a deduction test devised to measure the student's ability to judge good and poor arguments. This test was revised and called Test B. The reliabilities were not stable; they

---

<sup>74</sup> Thomas B. Edwards, "Measurement of Some Aspects of Critical Thinking." Unpublished Doctor's thesis, Department of Education, University of California, 1949. pp. 23-50.



ranged from .20 to .86. Test V was the judgment test discussed as Test C above. Test VI presented ten paragraphs, each of which was followed by three conclusions; one sound, one irrelevant, and one contradicted by the data. These were labeled A, B, and C and were presented in pairs. The student was instructed to choose the better of the pair. Test VII was similar to test VI, but the conclusions were all based upon the data. The student was instructed to choose the better of a pair of the conclusions. This test, upon revision, became Test D. The estimated reliabilities were .82 and .84. The correlation of this test with intelligence was .22.

Summary concerning tests on abilities involved in problem-solving. Considerable progress has been made in the testing of abilities involved in problem-solving in the three decades since Herring<sup>75</sup> published his test of scientific thinking. His pioneer work was of considerable interest because it was the first test of such a nature to be published and because he defined the kinds of behaviors which he associated with scientific thinking. Watson's<sup>76</sup> Test of Fairmindedness, though designed to measure prejudice, was a forerunner of most of the tests which have been devised to measure the ability to interpret data. In

---

<sup>75</sup> Herring, op. cit., pp. 535-558.

<sup>76</sup> Watson, op. cit., pp. 9-35.

addition, it was later modified by Watson and Glaser and became the highly successful Test of Critical Thinking. Watson's contribution was also significant in that he validated the test by curricular and statistical methods.

Another significant test of the mid-twenties was Zyve's<sup>77</sup> Stanford Scientific Aptitude Test, which purported to measure eleven scientific aptitudes. This test appears to have been the first test of this type and has been widely used. This test, also, was quite well validated. Downing's<sup>78</sup> test of scientific thinking was a distinct contribution because it was designed to measure many of the skills and safeguards of scientific thinking. The primary contribution of Weller<sup>79</sup> was the recognition of the distinction between the skills of scientific thinking and the scientific attitudes. One of the best of the attitudes tests was, "What Do You Think?", constructed by Noll,<sup>80</sup> who defined attitudes as habits of thinking. This test also has been widely used.

The tests devised for the Eight-Year Study<sup>81</sup> were

---

<sup>77</sup> Zyve, op. cit., pp. 525-546.

<sup>78</sup> Downing, op. cit., pp. 121-128.

<sup>79</sup> Weller, op. cit., pp. 90-97.

<sup>80</sup> Noll, op. cit., pp. 18-25.

<sup>81</sup> Smith and Tyler, op. cit., pp. 3-156.

noteworthy contributions to test construction because in the development of these tests the behaviors attending the major objectives were considered in detail, and because the abilities involved in critical thinking were recognized as major outcomes of secondary education. The Interpretation of Data tests devised for the Eight-Year Study have been used very extensively.

In the last decade the trend toward increased emphasis on the teaching of critical thinking has culminated in the production of a number of tests devised to test phases of this major objective. The Watson-Glaser Test of Critical Thinking,<sup>82</sup> previously referred to, was reported. Johnson<sup>83</sup> made a significant contribution in devising a test revolving around a single major problem. Teichman<sup>84</sup> and Alpern<sup>85</sup> devised interesting tests to appraise the abilities to draw conclusions from data and the ability to devise experiments, respectively.

An entirely new approach to the problem of measuring the ability to think scientifically was presented by Read<sup>86</sup> in his Non-verbal Test of Scientific Thinking. This test

---

<sup>82</sup> Glaser, op. cit., pp. 87-92.

<sup>83</sup> Johnson, op. cit., pp. 83-96.

<sup>84</sup> Teichman, op. cit., pp. 268-279.

<sup>85</sup> Alpern, op. cit., pp. 220-229.

<sup>86</sup> Read, op. cit., pp. 361-366.

was designed on the assumption that critical discrimination is the keynote of scientific thinking, and presents an interesting method of isolating this factor.

No attempt has been made in this summary to include mention of all of the tests and testing techniques which have been developed. Only the highlights in the measurement of problem-solving have been treated. It is, however, of interest to note, that tests have been devised for almost all educational levels from fourth grade through college, and that some tests have been devised without regard to subject matter areas, whereas, others have been designed for specific subjects.

#### RELATIONSHIP BETWEEN PROBLEM-SOLVING AND OTHER ABILITIES

Relation of intelligence to abilities involved in problem-solving. It is the opinion of a few investigators that the abilities involved in problem-solving are identical with intelligence. The majority of investigators seem to believe that there is a moderate to substantial relationship between intelligence and the abilities involved in problem-solving. A few, however, contend that the two abilities are almost completely unrelated.

Billings<sup>87</sup> has cited some evidence to support the

---

<sup>87</sup> Marion L. Billings, "Problem-solving in different fields of endeavor." American Journal of Psychology, 46:259-272, April, 1934.

viewpoint that problem-solving is a general intelligence factor. In an attempt to ascertain the nature of problem-solving, he presented his subjects with problems in eight different subject-matter areas. The subject matter necessary to the solution of the problems was taught prior to the administration of the tests. He obtained correlations ranging from .53 to .78 between the tests of reasoning in the various subject-matter areas. The average correlation was .67. Correlations between the tests of reasoning in the various fields and intelligence, as measured by the Army Alpha test, ranged from .42 to .59. Since he found a higher average correlation between the scores on reasoning in various fields than between reasoning in a particular field and information in that field, he inferred that problem-solving was an important part of Spearman's general factor of intelligence, if not intelligence itself.

It is interesting to note that Billings<sup>88</sup> attributed problem-solving to intelligence with correlations of from .42 to .59 between his test and an intelligence test, while other investigators obtaining similar correlations have not interpreted their data as indicating particularly high relationships between problem-solving ability and intelligence.

---

<sup>88</sup> Billings, loc. cit.

Zyve,<sup>89</sup> Sinclair and Tolman,<sup>90</sup> and Downing<sup>91</sup> seem to believe, however, that critical or scientific thinking is an innate characteristic. On the other hand, many investigators have shown that the ability to think scientifically can be taught. If this is true, problem-solving could not be identical with intelligence nor could it be an innate ability. A discussion of these alternate viewpoints follows.

Zyve,<sup>92</sup> who considered his test to be a measure of scientific aptitude, did not claim that the aptitude was intelligence itself. His data gave evidence that it was not intelligence, since he found a correlation of .44 to .51 between his test and intelligence as measured by the Thorndike intelligence test.

A study by Sinclair and Tolman<sup>93</sup> on the effect of scientific training on logical thinking showed that students in the science and engineering fields in college were superior to students in other fields in their ability to make inferences, as evidenced by the Inference test of the

---

<sup>89</sup> Zyve, op. cit., pp. 525-546.

<sup>90</sup> James H. Sinclair and Ruth S. Tolman, "An attempt to study the effect of scientific training upon prejudice and illogicality of thought." Journal of Educational Psychology, 24:362-370, May, 1933.

<sup>91</sup> Downing, op. cit., p. 128.

<sup>92</sup> Zyve, op. cit., pp. 525-546.

<sup>93</sup> Sinclair and Tolman, op. cit., pp. 362-370.

Watson test of Fairmindedness. The authors<sup>94</sup> suggested that this might mean that students who elect science and engineering show a tendency to superiority in this ability. This suggestion would lead one to believe that Sinclair and Tolman consider the ability to infer to be an innate ability. They report a correlation of .49 between scores on the Thorndike Intelligence test and scores on Watson's Inference test.

Downing<sup>95</sup> reported a correlation of .66 between his Test on Scientific Thinking and intelligence for students in the senior high school, and a correlation of .47 between these traits for students in the junior high school. He concluded that intelligence, as expressed by IQ, was different from the elements or safeguards of scientific thinking. It was his opinion that the elements of scientific thinking were due to inherited ability while the safeguards were the result of instruction. However, he does not present convincing evidence in support of this viewpoint. Strauss<sup>96</sup> found a correlation of .64 between scores on Downing's test and scores on the Otis Intelligence test. The 90 students used in this study were between the ages of 10 and 18.

---

<sup>94</sup> Sinclair and Tolman, loc. cit.

<sup>95</sup> Downing, op. cit., pp. 121-128.

<sup>96</sup> Strauss, op. cit., pp. 89-93.

Ter Keurst and Bugbee<sup>97</sup> administered their test on the scientific method to college freshmen and sophomores. They found correlations of .51 and .66, respectively, between the scores made by these groups on their test and the scores on the American Council on Education Psychological Examination. Since their test measured knowledge of the method of science rather than ability to use the scientific method, these correlations cannot justifiably be compared with the other correlations reported here.

Glaser<sup>98</sup> reported correlations ranging from .03 to .52 between intelligence, as measured by the Otis Mental Ability test, and the six tests which make up the Watson-Glaser Test of Critical Thinking. The correlation of scores on the entire critical thinking test with scores on the Otis Mental Ability test was .46 for the initial administration of the test and .48 for the final administration of the test.

Howell<sup>99</sup> attempted to discover the effect of debating on critical thinking. As a part of his study he correlated the composite scores on five of the six Watson-Glaser tests with intelligence quotients. He obtained a correlation of .63.

---

<sup>97</sup> Ter Keurst and Bugbee, op. cit., pp. 489-501.

<sup>98</sup> Glaser, op. cit., 142-147.

<sup>99</sup> William S. Howell, "The effect of high school debating on critical thinking." Speech Monographs, 10: 96-102, Annual, 1943.



In a study of the ability of ninth grade students to make conclusions, Teichman<sup>100</sup> found a correlation of .65 between the scores on his test and scores on a measure of mental ability. He found no significant relationship between intelligence and growth in the ability to make conclusions.

Higgins,<sup>101</sup> as a part of his study on the educability of adolescents in inductive ability, devised a test entitled Judge Conclusions. He found that the correlation between the scores on this test and scores on the Henmon-Nelson Test of Mental Ability was .54. Of particular interest, however, was his finding of a correlation of only .36 between his test and Thurstone's Induction Test. One would expect that his test, which he believed measured abilities involved in inductive reasoning, would have had a higher correlation with a test which purported to measure the inductive factor of intelligence than with a general intelligence test, such as the Henmon-Nelson Test of Mental Ability.

Weisman,<sup>102</sup> in her study of factors related to the ability to interpret data, reported correlations of .64 to

---

<sup>100</sup> Teichman, op. cit., pp. 268-279.

<sup>101</sup> Higgins, op. cit., p. 40.

<sup>102</sup> Leah L. Weisman, "Some Factors Related to the Ability to Interpret Data in Biological Science." Unpublished Doctor's thesis, Department of Education, University of Chicago, 1946. p. 91.

.69 between intelligence as measured by the Henmon-Nelson Test of Mental Ability and ability to interpret data as measured by the Progressive Education Association Interpretation of Data test.

The studies considered thus far have all given evidence of a moderate to substantial relationship between intelligence and problem-solving abilities. Two studies, utilizing the technique of partial correlations, have shown that the true relationship between intelligence and problem-solving is probably not shown by simple correlations. In a study devised to investigate the relationship between ability to recall and ability to reason, Smith<sup>103</sup> found a correlation of .58 between ability to reason and IQ. When ability to recall was held constant, by means of a partial correlation, this coefficient of correlation between ability to reason and IQ was reduced to .23. Alpern,<sup>104</sup> in his study on the ability of students to test hypotheses, found a correlation of .53 between intelligence and ability to test hypotheses. However, by holding reading grade and chronological age constant by the use of a partial correlation, he found the correlation was reduced to .11.

---

<sup>103</sup> Victor C. Smith, "A study of the degree of relationship existing between ability to recall and two measures of ability to reason." Science Education, 30:88-90, March, 1946.

<sup>104</sup> Alpern, op. cit., pp. 222-223.

Somewhat lower correlations between intelligence and abilities involved in critical thinking have been reported in a number of studies. Hoff<sup>105</sup> reported a correlation of .36 between intelligence as measured by the American Council on Education Psychological examination and his test for scientific attitudes. Noll<sup>106</sup> found moderate positive correlations, ranging from .30 to .41 between IQs and scores on preliminary forms of his test, "What Do You Think." These correlations, he believed, indicated that his test measured factors other than intelligence or native ability of the eighth to twelfth grade students to whom he administered the tests.

Bedell,<sup>107</sup> in a study on the relation between the ability to infer and the ability to recall, found low positive correlations between intelligence of junior and senior high school students and their ability to infer. However, his data revealed that the lowest quarter of the group, in terms of scores on the intelligence test, scored scarcely better than chance on the inference test. He concluded, tentatively, that a certain degree of intelligence is essential to problem-solving ability.

---

<sup>105</sup> Hoff, op. cit., pp. 28-35.

<sup>106</sup> Noll, op. cit., p. 24.

<sup>107</sup> Ralph C. Bedell, "The Relationship Between the Ability to Infer in Specific Learning Situations." Unpublished Doctor's thesis, Department of Education, University of Missouri, 1934. pp. 36-37.

Johnson<sup>108</sup> correlated scores made on her test devised to measure reflective thinking with mental alertness, as measured by the Ohio Psychological examination. She reported a coefficient of correlation of .40 for a group of 84 college students. She believed that the data revealed that those aspects of reflective thinking measured by her test may depend on college level intelligence, but that other variables were more significant.

Furst,<sup>109</sup> in a study of changes evoked in two years of general education, gave a series of tests to measure, among other things, changes in the ability to think critically. As a part of his study, he correlated the scores made on the portions of his test which measured critical thinking with intelligence as measured by the American Council on Education Psychological examination. He found that 80 percent of these correlations were below .40. He asserted that his data indicated that the various tests of critical thinking measured characteristics of student's behavior which were not highly related to measures of scholastic aptitude. He believed that, at the secondary school level and the lower college level, students with relatively low scholastic aptitude may

---

<sup>108</sup> Johnson, op. cit., pp. 83-96.

<sup>109</sup> Edward J. Furst, "Changes in Organization of Various Abilities and Skills after Two Years of General Education at the Secondary-School Level." Unpublished Doctor's thesis, Department of Education, University of Chicago, 1948. p. 155.

be able to perform as well as those with high scholastic aptitude on tests of critical thinking.

Dunning<sup>110</sup> studied the relationship of the ability to interpret data, as measured by his test, to factors of intelligence. As a measure of the factors of intelligence he used a battery of Thurstone's Primary Mental Abilities tests. He found correlations of from .04 to .24 between the various factors of intelligence as measured by this test and the scores on the interpretation of data portion of his Test of Scientific Thinking. He concluded that the ability to interpret data was a different ability than any of the factors of intelligence.

Read<sup>111</sup> reported a correlation of .39 between intelligence and his non-verbal test of the ability to use the scientific method. Edwards<sup>112</sup> found correlations ranging from .00 to .22 between measures of intelligence and his four tests which were designed to measure (1) induction, (2) deduction, (3) judging opinions, and (4) judging conclusions.

---

<sup>110</sup> Gordon M. Dunning, "The construction and validation of a test to measure certain aspects of scientific thinking in the area of first year college physics." Science Education, 33:221-235, April, 1949.

<sup>111</sup> Read, op. cit., pp. 261-266.

<sup>112</sup> Edwards, op. cit., pp. 80-85.

Fleming<sup>113</sup> studied the outcomes of a course in biology at the college level. One of the purposes of his investigation was to measure growth in understanding of the elements of the scientific method. As a part of this study he correlated the scores made on his test of scientific thinking with intelligence. He reported a coefficient of correlation of .34.

Summary of studies concerning the relation of intelligence to problem-solving. There is no substantial agreement among investigators concerning the relationship of problem-solving to intelligence. A number of investigator's correlations ranged from .40 to .69, indicating a fairly substantial relationship between intelligence and problem-solving abilities. Billings<sup>114</sup> interpreted such correlations as indicating that problem-solving ability is a general factor, if not intelligence itself, whereas other investigators made no such claim. On the other hand, however, some investigators have found correlations ranging from .00 to .40, indicating no relationship to moderate relationship between these characteristics. Evidence obtained by the use of partial correlations indicated that other factors, such as memory and reading ability may account for some of the relatively high correlations.

---

<sup>113</sup> Fleming, op. cit., p. 185.

<sup>114</sup> Billings, op. cit., pp. 259-272.

Although many of the correlations show a moderate to substantial relationship between intelligence and the abilities involved in problem-solving, these correlations are not as high as correlations between scores on intelligence tests and achievement tests over information previously learned. Stroud<sup>115</sup> has stated that correlations between scores on achievement batteries and intelligence tests are of the magnitude of .8, and Kelley<sup>116</sup> claimed that there was a 90 percent overlapping between a general intelligence test and a general achievement test. These findings seem to indicate that there is somewhat less relationship between intelligence and ability to think scientifically than between intelligence and general academic achievement.

Zyve,<sup>117</sup> Downing,<sup>118</sup> and Sinclair and Tolman<sup>119</sup> support the viewpoint that the ability to think critically is an innate characteristic. If this is true, no appreciable improvement in scores on thinking tests as a result of instruction would be anticipated. Evidence to the contrary

---

<sup>115</sup> James B. Stroud, Psychology in Education. New York: Longmans, Green and Company. 1946. pp. 338-339.

<sup>116</sup> Truman L. Kelley, Interpretation of Educational Measurements. Yonkers-on-Hudson: World Book Company. 1927. pp. 363.

<sup>117</sup> Zyve, op. cit., pp. 525-546.

<sup>118</sup> Downing, op. cit., pp. 121-123.

<sup>119</sup> Sinclair and Tolman, op. cit., pp. 262-270.

is presented in the discussion which follows.

Educability in problem-solving. Related to the problem of the relationship of intelligence to abilities involved in critical thinking, is the problem of educability in the thinking process. If abilities involved in critical thinking were primarily due to intelligence as suggested by Billings, there should be little, if any, improvement in the ability with training. The evidence seems to indicate that these abilities can be improved if they become specific objectives of instruction. On the contrary, there is no evidence to indicate that they are a necessary by-product of the study of science. As indicated by Noll,<sup>120</sup> the attainment of these objectives will come when they are taught; that is, when the emphasis of teaching is upon learning to think rather than on memorization of facts.

There is considerable evidence to show that skills of the scientific method can be taught effectively to students of all grade levels. Weller<sup>121</sup> found a significant difference between two equated groups of sixth grade students; one group received specific instruction in both scientific attitudes and skills of scientific thinking, while the other received no special training. She concluded that growth in

---

<sup>120</sup> Victor H. Noll, "Teaching the habits of scientific thinking." Teachers College Record, 35:202-212, December, 1933.

<sup>121</sup> Weller, op. cit., pp. 90-97.



both attitudes and skills could be stimulated if they were specific objectives of instruction. Arnold,<sup>122</sup> in a study of fifth and sixth grade students, also concluded that critical thinking can be taught in the elementary school.

Grener and Rath<sup>123</sup> found significant gains in the ability to think critically in a group of third grade pupils after a five month period of teaching for critical thinking.

Curtis<sup>124</sup> and Daily<sup>125</sup> both found that junior high school pupils benefited from direct instruction in critical thinking.

Blair and Goodson<sup>126</sup> conducted an experiment which showed that ninth grade students receiving instruction in scientific thinking improved more on Noll's<sup>127</sup> "What Do You Think" test than did the two groups which did not receive this special instruction. One of the control groups<sup>128</sup>

<sup>122</sup> Dwight Arnold, "Testing Ability to use data in the fifth and sixth grades." Educational Research Bulletin, 17:255-259, December, 1937.

<sup>123</sup> Norma Grener and Louis E. Rath, "Thinking in third grade." Educational Research Bulletin, 24:38-42, February, 1945.

<sup>124</sup> Curtis, op. cit., p. 78.

<sup>125</sup> Daily, op. cit., p. 81.

<sup>126</sup> Glenn M. Blair and Max R. Goodson, "Development of scientific thought in general science." School Review, 47:696-700, November, 1939.

<sup>127</sup> Noll, Habits of Scientific Thinking, op. cit., pp. 27.

<sup>128</sup> Blair and Goodson, op. cit., pp. 696-700.

received no science instruction, while the other control group received science instruction by the usual methods. The means for all three groups were higher on the post-test than on the pre-test. The comparison of means for the two control groups showed no significant difference which seems to support Downing's<sup>129</sup> viewpoint that science instruction does not necessarily produce growth in ability to think scientifically.

Teichman<sup>130</sup> investigated the ability of ninth grade students to draw conclusions. Twelve groups, designated as controls, were taught the regular course in science. Eight groups were given additional training in the drawing of conclusions. He found that although both groups made gains in these abilities, the experimentals made significantly greater gains.

Higgins<sup>131</sup> studied the educability of adolescents in inductive ability. He reported that the gains of students receiving special instruction in problem-solving in a course in high school biology were meaningfully greater than the gains of other students taking biology but not receiving special instruction in problem-solving.

---

<sup>129</sup> Downing, op. cit., pp. 121-128.

<sup>130</sup> Teichman, op. cit., pp. 268-279.

<sup>131</sup> Conwell D. Higgins, "The educability of adolescents in inductive ability." Science Education, 29:82-85, March, 1945.

Neuhof<sup>132</sup> found that students taking high school chemistry improved markedly in their ability to interpret data, as measured by the Progressive Education Association Interpretation of Data tests, after training in the interpretation of data. No control group was employed in this study. Gains in scores were not limited to the better students. He concluded that definitely measurable results could be achieved in the teaching of such complex mental processes as the interpretation of data.

Weisman<sup>133</sup> investigated the development of skills of scientific thinking in high school biology. Six classes taught by the investigator using problem-solving techniques were compared with six classes taught by teachers who believe that the ability to think scientifically could be taught without special instruction. Weisman found her experimental groups gained significantly more than the controls on the Progressive Education Association Interpretation of Data tests. There was also a significant gain on several of the Watson-Glaser Tests of Critical Thinking. Although these results are consistent with results of many other studies, Mallison<sup>134</sup> criticized the implication of the finding because

---

<sup>132</sup> Mark Neuhof, "Integrated interpretation of data tests." Science Education, 26:21-26, January, 1942.

<sup>133</sup> Weisman, op. cit., pp. 77-83.

<sup>134</sup> George G. Mallison, "The implications of recent research in the teaching of science at the secondary-school level." Journal of Educational Research, 43:321-342, January, 1950.

the study failed to take into account the fact that the investigator may have been a superior teacher.

Glaser<sup>135</sup> utilized four control and four experimental classes in twelfth grade English to measure changes in ability to think critically. The experimental classes were given instruction to stimulate critical thinking. Glaser found that the average gains on the battery of critical thinking tests of the four experimental classes, after ten weeks of instruction, were significantly greater than the average gains of the control classes. This study is especially significant in that it included a follow-up study. The students were tested again six months after the experimental period. The growth in ability to think scientifically had been retained. Glaser predicted that some aspects of the growth would probably be retained more or less permanently, and would afford a basis for further growth in the ability to think critically.

A few studies have been reported on teachability of the skills involved in scientific thinking at the college level. Teller<sup>136</sup> used an experimental and a control group of students taking a course in the history of education. Both groups had classes five days a week, but one class

---

<sup>135</sup> Glaser, op. cit., pp. 131-140.

<sup>136</sup> James D. Teller, "Improving ability to interpret educational data." Educational Research Bulletin, 19:363-371, September, 1940.

period each week was devoted to the interpretation of historical data in the experimental section. Teller found that the experimental group showed greater improvement in the ability to interpret data as measured by a test constructed to appraise the ability to interpret historical data.

Tyler<sup>137</sup> reported a study on remedial instruction for students enrolled in a course in freshman zoology. Students who received remedial instruction in problem-solving techniques gained significantly more than those without the remedial instruction. Students in this study were matched on the basis of intelligence, pre-test scores, sex, and instructor.

Fleming<sup>138</sup> reported a study to measure certain outcomes of a course in biological science. One of the outcomes appraised was the ability to think scientifically. He equated two groups of students, one taking no science, the other taking biological science. He found that, although both groups made gains in the ability to think scientifically, those taking the science course made significantly greater gains.

Thelen<sup>139</sup> made a study of the effect of instruction

---

<sup>137</sup> Ralph W. Tyler, Service Studies in Higher Education. Columbus, Ohio: The Ohio State University. 1932. pp. 119-122.

<sup>138</sup> Fleming, op. cit., pp. 172-179.

<sup>139</sup> Thelen, op. cit., pp. 234-261.

planned to produce growth in the ability to think scientifically. The experiment was conducted with students in a course in freshman chemistry. The control groups were taught by traditional laboratory methods, whereas the experimental groups were given opportunities to participate in inductive thinking as often as was feasible. Thelen's test on experimental procedures and the Progressive Education Association Interpretation of Data test were used to evaluate these abilities. Using the technique of analysis of covariance, he found that the experimental groups were superior to the controls. However, the gains were not great in terms of percent gains.

Bond,<sup>140</sup> in a study similar to Thelen's found superiority in an experimental group. The subject-matter area of Bond's study was a unit on genetics in a course in college biology.

Barnard<sup>141</sup> compared the relative effectiveness of the lecture-demonstration method with the problem-solving method in the teaching of a course in college science. The evaluation instruments included a test on the ability to solve

---

<sup>140</sup> Austin D. M. Bond, An Experiment in the Teaching of Genetics with Special Reference to the Objectives of General Education. Contributions to Education, No. 797. New York: Bureau of Publications, Teachers College, Columbia University. 1940. pp. 77-79.

<sup>141</sup> J. Darrell Barnard, "The Lecture-demonstration vs problem-solving method of teaching a college science course." Science Education, 26:121-132, October, 1942.

problems. The groups used were equated on the basis of pre-tests and scores on psychological examinations. He found that the problem-solving method produced significantly greater gains on the tests designed to measure problem-solving abilities.

Summary of studies on educability in problem solving.

The evidence presented in this portion of the review of literature seems to indicate that abilities involved in critical thinking are not to any considerable extent a by-product of the teaching of science. The evidence also lends credence to the hypothesis that critical thinking can be taught providing it is a specific objective of instruction. However, the evidence is still fragmentary and the conclusion is tentative.

Relation of reading ability to the abilities involved in problem-solving. There is considerable evidence to show that there is a relationship between reading ability and the ability to think critically. An interesting point in this regard is the fact that Buros<sup>142</sup> placed the Progressive Education Association Interpretation of Data tests among his list of reading tests in the 1940 Mental Measurement Yearbook.

Grim<sup>143</sup> found correlations ranging from .51 to .66

---

<sup>142</sup> Oscar K. Buros, The Nineteen-Forty Mental Measurement Yearbook. Highland Park, N.J.: The Mental Measurement Yearbook. 1941. pp. 346-347.

<sup>143</sup> Paul R. Grim, "Interpretation of data and reading ability in social studies." Educational Research Bulletin. 19:372-374, September, 1940.

between scores on Progressive Education Association Interpretation of Data tests and scores on reading tests among junior high school students. Weisman<sup>144</sup> also used the Progressive Education Association Interpretation of Data test in her study on factors related to the ability to interpret data among high school students. She found correlations between scores on this test and scores on the Iowa Silent Reading test to range from .57 to .65. A partial correlation between scores on the reading test and scores on the interpretation of data test with IQ held constant was .34. Dunning<sup>145</sup> compared scores on his interpretation of data test in physics, designed for college freshmen, with scores on a reading test. He reported a correlation of .36.

Glaser<sup>146</sup> reported correlations of .32 and .36 between the composite score on the Watson-Glaser battery of tests and scores on the Nelson-Denny reading test. Correlation of scores on the reading test and scores on the six individual tests of the Watson-Glaser battery ranged from -.06 for the generalization test to .55 for the inference test. It is of interest to note that there is a relatively high correlation between reading ability and an ability to judge the degree of truth or falsity of statements. Glaser<sup>147</sup> found

---

<sup>144</sup> Weisman, op. cit., pp. 97-98.

<sup>145</sup> Dunning, op. cit., p. 232.

<sup>146</sup> Glaser, op. cit., pp. 142-147.

<sup>147</sup> Ibid., pp. 166-167.



higher correlations between scores on his test and a test of Reading Comprehension. These correlations ranged from .36 to .77 for the individual tests and from .77 to .82 for his battery of tests.

Ter Keurst and Bugbee<sup>148</sup> obtained correlations of .57 and .59 between scores on their test on scientific method and scores on the Nelson-Denny reading test. As previously mentioned, this test of scientific method seems to measure knowledge of steps and attitudes rather than behaviors. On this basis one might expect rather high correlation between reading ability and scores on this test.

Teichman,<sup>149</sup> in studying the ability of ninth grade students to draw conclusions, found a correlation of .61 between this ability as measured by his test and reading ability. Alpern<sup>150</sup> found similar correlations between his test on ability to test hypotheses and reading grade in high school pupils. He reported a correlation of .57. However, Alpern found that by holding IQ constant by means of a partial correlation, this correlation was reduced to .36.

Hoff<sup>151</sup> found low correlations between his test and reading ability. He reported a correlation of .19 between

---

<sup>148</sup> Ter Keurst and Bugbee, op. cit., pp. 489-501.

<sup>149</sup> Teichman, op. cit., pp. 268-279.

<sup>150</sup> Alpern, op. cit., pp. 220-229.

<sup>151</sup> Hoff, op. cit., pp. 28-35.

scores on his test and scores on the comprehension portion of the American Council on Education Reading test. The correlation between scores on his test and speed of reading scores on the American Council on Education Reading test was .09.

Summary of studies concerning the relation of reading to problem-solving. The evidence presented seems to indicate that reading ability and ability to think scientifically are to some degree related. Interpretation of data tests and other tests measuring abilities involved in scientific thinking are to a substantial degree dependent upon reading ability. On the other hand, scores on attitude tests did not seem to depend to any marked extent on reading ability.

Relation of factual information to the abilities involved in problem-solving. According to Wood and Beers,<sup>152</sup> thinking and thinking ability are not under the control of teaching except as thinking is influenced by knowledge. This statement seems to imply that general intelligence and knowledge of facts should account for all of the variability in scores on thinking tests. The evidence for this point of view is somewhat contradictory as may be seen in the following discussion.

---

<sup>152</sup> Ben D. Wood and F. S. Beers, "Knowledge versus thinking." Teachers College Record, 37:487-499, March, 1936.

Bedell<sup>153</sup> planned a study to determine the relationship between the ability to recall and the ability to infer. Thirty paragraphs containing facts from which the student could infer principles were given. Two sets of test items were constructed; one to measure knowledge of facts, one to measure the ability to make inferences. These tests were administered to 324 students in junior and senior high schools. Bedell found that the ability to recall and the ability to infer were different but not completely unrelated. According to his findings the ability to infer was a more difficult process than the ability to recall.

Billings,<sup>154</sup> in studying problem-solving in different areas found higher correlations between the ability to solve problems in different areas than between ability to solve problems and information in the same area. Correlations between scores in problem-solving in various fields ranged from .53 to .78. The average correlation was .67. The average correlation between scores on information tests and problem-solving tests in the same field was .45. He concluded that those who solved the problems know the material, but that not all who knew the material could solve the problems.

Smith<sup>155</sup> found high correlations between the ability

---

<sup>153</sup> Bedell, op. cit., pp. 10-50.

<sup>154</sup> Billings, op. cit., pp. 259-272.

<sup>155</sup> Smith, op. cit., pp. 88-90.

to reason and knowledge of facts. The correlation he obtained was .77. The reduction in this correlation was slight when IQ was held constant by means of a partial correlation. The partial correlation was .65. He concluded that the ability to recall information and the ability to see relationships seemed to be two products of the same learning process.

Dunning<sup>156</sup> reported a correlation of .56 between his interpretation of data test for a physics course for college freshmen and a factual information test covering the same topics. Since his correlation of .56 indicated a 38 percent overlapping between the interpretation of data test and the factual test, he concluded that knowledge of factual information was no guarantee of ability to use the information in the solving of problems.

Fleming,<sup>157</sup> as a part of his study on outcomes in a course in biology at the college level, reported a correlation of .57 between the test he used to measure the ability to think scientifically and the test he used to measure knowledge of facts.

Weisman,<sup>158</sup> in a study of factors related to the ability to infer, reported a correlation of .63 between scores

---

<sup>156</sup> Dunning, op. cit., p. 232.

<sup>157</sup> Fleming, op. cit., pp. 186-187.

<sup>158</sup> Weisman, op. cit., pp. 104-105.

on the Progressive Education Association Interpretation of Data test and scores made on the Cooperative Biology test. She found, however, that there was little relationship between scores on the interpretation of data test and gain in knowledge of biology, or between gain in ability to interpret data and knowledge of facts.

Read<sup>159</sup> found a correlation of .53 between scores on his non-verbal test of the ability to use the scientific method and scores made on the Cooperative General Science test.

In a course in elementary biology, Tyler<sup>160</sup> found a correlation of .41 between scores on an information test and scores on a test measuring ability to interpret data. He reported a correlation of .46 between scores on the information test and a test designed to measure the ability to plan experiments to test hypotheses, and a correlation of .35 between knowledge of technical terms and ability to draw inferences. In another study of college students taking various subjects, he<sup>161</sup> found correlations ranging from .20 to .53 between scores on tests of recall and scores on

---

<sup>159</sup> Read, op. cit., pp. 361-366.

<sup>160</sup> Ralph W. Tyler, "Measuring the results of college instruction." Educational Research Bulletin, 11:253-260, May, 1932.

<sup>161</sup> Ralph W. Tyler, in Charles H. Judd, Education as Cultivation of the Higher Mental Processes. New York: The Macmillan Company. 1936. p. 14.

tests requiring students to draw inferences and concluded that there was little relationship between these two abilities.

Summary of studies concerning the relation of knowledge of facts to problem-solving abilities. The evidence indicates that there is a moderate positive correlation between the abilities involved in problem-solving and the knowledge of facts. These findings seem, in general, to support the conclusion that facts are essential to thought and to problem-solving, but that knowledge of facts does not guarantee an ability to use the facts in the solution of a problem.

Summary of research related to the problem. An attempt has been made in this chapter to show how the descriptive analysis of the steps of scientific thinking is related to measurement of the ability to think scientifically, and how the development of tests has influenced educational research on the ability to think scientifically. Early work in the descriptive analysis was done by philosophers and individual scientists, but no systematic evaluation of the steps involved in scientific thinking was attempted until about twenty-five years ago. Since that time important contributions to an understanding of the nature of scientific thinking have been made by various workers. Such an understanding is of special importance to the measurement of

ability to think scientifically because the steps or elements of scientific thinking provide specific objectives to be tested, and because the steps offer suggestions of the types of behaviors which attend or which represent scientific thinking.

The recognition of the ability to think scientifically as a major objective of education stimulated the construction of tests to appraise various aspects of this ability. This testing movement, while slow at first, has resulted in the production of a number of tests which are quite reliable and which seem to have considerable validity. A variety of techniques have been evolved to measure the abilities involved in scientific thinking. Many of the techniques appear to be useful methods of obtaining evidence of the abilities.

The development of instruments to measure scientific thinking has led to studies of the relationship of this ability to various other traits such as, intelligence and reading ability, and to the knowledge of facts. The evidence presented supports the inference that there is a direct relationship between the ability to think scientifically and the above mentioned traits. However, most investigators are of the opinion that these factors do not account for all of the variability in scores on tests designed to measure ability involved in problem-solving.

One of the most stimulating findings of the investigators into the nature of problem-solving is that it can, apparently, be taught; particularly if it is a specific objective of instruction. The bulk of evidence supports this view.



## CHAPTER III

### GENERAL PROCEDURES INVOLVED IN THE DEVELOPMENT OF THE TEST

The purpose of this chapter is to describe: (1) the manner in which the test was developed, (2) the methods used in the construction of the test items, (3) the nature of the groups to which the test was administered in its various stages of development, (4) the methods used in the statistical analysis of the test, and (5) the methods used in the validation of the test.

The general procedures followed in the development of the test to measure the ability to think scientifically were similar to those used by Smith and Tyler<sup>1</sup> in the development of the tests used in evaluating the results of the Eight-Year Study. Several steps in the process and a detailed description of the procedure within each step as modified for its use in the present study are given below.

The first four steps were: (1) the setting up of the objectives, (2) the definition of each of these objectives in terms of desired behavior, (3) the identification of situations in which students could be expected to display these behaviors, and (4) the writing of items to evaluate

---

<sup>1</sup> Eugene R. Smith, Ralph W. Tyler and the Evaluation Staff, Appraising and Recording Student Progress. New York: Harper and Brothers. 1942. pp. 15-28.

the behaviors. The fifth step was the tryout of the items constructed, the analysis of these items, and the incorporation of the best items into a test to measure the ability to think scientifically. The sixth step was the administration and analysis of this test. The seventh step was the validation of the test.

Detailed discussions of methods used in the construction, analysis and validation of the test are reserved for the chapters which deal with these aspects of the problem because it was felt that these discussions would be more meaningful when presented with the materials with which they were used. A complete treatment of the first four steps is presented in Chapter IV. Chapter V is devoted to a detailed discussion of steps five and six. Chapter VI deals with the validation of the test, step seven.

The first step in the construction of the test designed to appraise the ability to think scientifically was the formulation of the educational objectives to be measured. The formulation of the objectives involved a consideration of the elements involved in scientific thinking as discussed in Chapter II, and a consideration of the objectives of teaching implied by each of these elements.

The second step was the definition of each of these objectives into terms of desired behavior. As Tyler<sup>2</sup> has

---

<sup>2</sup> Ralph W. Tyler, Constructing Achievement Tests. Columbus, Ohio: Ohio State University. 1934. pp. 4-23.

stressed, this step is one of the crucial ones of test construction since objectives are usually stated in rather broad general terms. For example: the ability to interpret data is an oft-stated objective of science teaching. But what are the specific things that a person does when he interprets data? What are the kinds of errors made by persons who do not consistently achieve this objective? In order to determine what these behaviors are a study must be made of the types of reactions made by persons who are competent in this objective.

Sources of these behaviors were (1) the major and minor elements involved in scientific thinking, (2) literature on test construction, especially on tests devised to measure various aspects of scientific thinking, (3) committee reports on behaviors involved in scientific thinking, (4) reports of research on behaviors of persons doing scientific research, and (5) interviews with teachers of science who are attempting to teach scientific thinking.

The third step was the identification of situations in which students could be expected to display the types of behaviors identified in step two. It was deemed advisable to select materials, which would be of some interest to the students, which dealt with biological subject matter free of technical terms, and which would be comprehensible to students who had had no previous experience with biological subject matter. Technical journals, popular journals and

textbooks were examined for situations which could be utilized in the construction of test items.

The fourth step involved the selection and trial of promising methods of measuring behaviors which would give evidence of the attainment of the objectives. This step included the writing of the items and the organization of tryout tests. It is customary to construct two to five times as many items as used in the final form of the test so that poor items may be eliminated. For this reason a series of nine tryout tests was constructed. Each of these tests was designed to measure a limited number of the behaviors involved in scientific thinking.

The tests were designed so that they could be scored on International Business Machine answer sheets. The five choice answer sheet was selected as the most appropriate for the purpose of this test. The detailed discussion of the construction of the tryout tests and examples of items from each of them will be presented in Chapter IV.

A total of 637 items was constructed for the nine tryout tests. They were given to four members of the department of Biological Science at Michigan State College and to one expert in the field of testing in biological science for criticisms and suggestions.

The fifth step was the administration of the tryout tests, the determination of the difficulty and validity of

the items, and the selection of the best items. The tryout tests were administered to a group of 168 students taking the third term of the three-term sequence of Biological Science at Michigan State College during the spring term of 1950. Only students for whom comparable psychological and reading examination scores were available were used in this testing. For this reason only students who had taken the examinations given to entering freshmen by the Board of Examiners in the fall of 1949 were admitted to the six sections which had been designated as experimental sections for tryout tests. However, six of the 168 students actually enrolled in these sections were not freshmen who had entered Michigan State College in the fall of 1949. These students had been pre-registered in one of the experimental sections by the department of Engineering. Consequently they could not be transferred to other sections. The scores of these students on the tests were used in the calculation of means, standard deviations and reliabilities. The papers of these students were also used in the calculation of item difficulties and item validities but their scores were not utilized in the computation of correlations between test scores and scores on intelligence tests, reading tests, and factual tests.

Of the 168 students to whom the tryout tests were given 83 were males and 85 were females. The age range of this group at the beginning of the spring quarter was from 17 years to 25 years; the mean age was 18.76 years.

The tryout tests were given during each alternate laboratory period during the term. The laboratory period was one hour and fifty minutes in length. Students were permitted to work at their own rate of speed on these tests and all students were allowed to finish all of the items on all of the tests. Some students finished as many as three of the tryout tests during one period while others completed only one or two per period. The students were instructed to answer all items even if it was necessary to guess. All of the tests were scored on the basis of the total number of correct answers and no correction for chance was used in the scoring.

As previously mentioned, 162 of the students completing the testing program had entered Michigan State College in the fall of 1949. At that time they had been given the 1949 edition of the American Council on Education Psychological Examination, which purports to measure the linguistic and quantitative factors of intelligence. A composite score, referred to as the total psychological score, is obtained as well as a score on the linguistic portion of the test and a score on the quantitative portion of the test. Form Y of the American Council on Education Reading Comprehension Test was administered at the same time. This test yields a total reading score, a vocabulary score, a speed of reading comprehension score and a level of reading comprehension score.

At the completion of the year course in Biological Science a comprehensive examination covering the year's work in biology is given to the students. This examination is prepared by the Board of Examiners of Michigan State College. The score obtained by the student on this examination determines the mark which he receives for the entire year's work. The comprehensive examination scores were obtained for each of the 168 students to whom the tryout tests were administered. In addition, the comprehensive examination papers of these students were rescored on the basis of items which were purely factual and items involving the ability to think scientifically. The latter items differ from the items of the tryout tests in that they involve a knowledge of biological facts and principles. Of a total of 300 items in the comprehensive examination, 53 were purely factual while 247 required some use of skills involved in scientific thinking.

Although the student's mark in biological science is determined entirely by his performance on the comprehensive examination, his progress through the three terms of the course is dependent upon the kind of work he does during the year. The work accomplished is reflected on the term-end examinations which are constructed and directed by a committee composed of members of the department of Biological Science. The scores made by each of the 168 students on their term-end examinations for the first and second terms

of the course were obtained.

The means and standard deviations were calculated for each of the tryout tests and for the entire battery of tests considered as a single test. The reliabilities of each of the tryout tests were calculated by correlating the scores on the odd-numbered items with the scores on the even-numbered items. These correlations gave reliabilities of a test half as long as the actual tests. The corrected reliabilities of each of the tryout tests were estimated by means of the Spearman-Brown prophecy formula. The reliability of the test battery was calculated by one of the Kuder-Richardson formulas.<sup>3</sup> The Kuder-Richardson formulas were designed to overcome the disadvantages of test-retest, equivalent forms, and split-half methods. Adkins<sup>4</sup> states that they are superior to other methods of determining the reliabilities of tests. The formula used in this study required only the number of items of the test, the mean of the test, and the standard deviation of the test. It is well to note that there are certain assumptions upon which this method rests. These assumptions are (1) that the test measures only one factor, (2) that the intercorrelation of all items are equal, and (3) that the items are equal in difficulty. If the

---

<sup>3</sup> Dorothy C. Adkins, Construction and Analysis of Achievement Tests. Washington: U. S. Government Printing Office. 1947. p. 153-154.

<sup>4</sup> Loc. cit.



assumptions are not met, the value obtained is an underestimate of the reliability. The value obtained represents the minimum reliability of the test for this group.

Item analysis is the analyzing of each item of a test to determine its validity and difficulty. Item analysis data were obtained for all items of all of the tryout tests. Item validity may be defined as a measure of the item's correlation with a criterion.<sup>5</sup> The purpose of determining the validity of the items is to identify items which discriminate well. Items difficulty is usually expressed as the percent of persons answering the item correctly. Since items answered correctly by almost all of the students or by almost none of the students cannot have any functional value in an achievement test inasmuch as they do not serve to discriminate between students, it is generally considered desirable to eliminate them. A detailed discussion of the methods used in the validation of the test items and in the calculation of the item difficulties will be presented in Chapter V.

The scores on each tryout test were correlated with the scores on each of the other tryout tests. This was done to determine whether a large degree of overlapping existed between the tests and to determine whether any tests might be eliminated in the construction of the single test used to measure the ability to think scientifically. The scores on

---

<sup>5</sup> Ibid., p. 180.

each of the tryout tests were also correlated with the quantitative score and the linguistic score on the American Council on Education Psychological Examination and with the total score on the reading test. These correlations were in reality measures of some phase of intelligence or of reading ability.

The purpose of administering the tryout tests was to identify good items to be used in the construction of a test to measure the ability to think scientifically. The tryout tests went through two revisions. The first revision resulted in a test, referred to as Test I, consisting of 150 items. This test was too long to be administered in the hour and fifty minute laboratory period, therefore twenty-five of the poorer items were eliminated from it. This final form of the test consisting of 125 items, is hereafter referred to as Test IA. Both Test I and Test IA have been called, The Ability to Think Scientifically. In the construction of Test I it was necessary, in most cases, to select blocks of items from the tryout tests rather than individual items since items were presented in blocks centering around a particular problem of experiment. The best blocks of items from each of the tryout tests, as determined by item analysis, were selected for inclusion in Test I. Poor items, as identified in the same manner, were eliminated from these blocks of items unless they were necessary to the development of the concept developed within the block of items. A total of 150

items were chosen to comprise Test I.

The sixth step involved the administration of Test I, the determination of the mean, standard deviation, and reliability of the test, and the analysis of the individual items. Test IA, the final form of the test, was constructed from Test I by the elimination of 25 of the poorer items. The sixth step also included the administration and statistical analysis of this final form of the test.

Test I was given in May, 1950, to 500 students who had completed the three-term sequence of Biological Science. This group has not previously been mentioned in this study. Of this group 291 were males and 209 were females. The age range was from 17 years to 37 years. The mean age was 20.04 years. Two hundred and sixty-four were freshmen who had entered Michigan State College in the fall term of 1949 and who had taken the 1949 edition of the American Council on Education Psychological Examination and Form Y of the American Council on Education Reading Comprehension Test at that time. The remaining students were either freshmen who had taken entrance examinations during the summer of 1949 or they were sophomores, juniors, or seniors. These students had all been given alternate forms of the American Council on Education Psychological Examination and the American Council on Education Reading Comprehension Test. Correlations of scores on Test I with scores on psychological examinations and with scores on the reading test were therefore based on

the score of these 264 students who had taken the forms of the latter tests given in the fall of 1949. This was done because it could not be assumed that scores on the various forms of these tests were directly comparable, and because raw scores were not available for any of the examinations given prior to the fall of 1949. Prior to this time only percentiles had been available.

The mean and the standard deviation were calculated for the group which completed Test I in the spring of 1950. An estimate of the reliability of the test for this group was determined by correlating the scores on the odd-numbered items with the scores on the even-numbered items. These correlations were adjusted for the total test by means of the Spearman-Brown formula. A second method used to determine the reliability of the test was the Kuder-Richardson formula. This calculation was done to compare the reliability obtained by the split-half method with a method which gives a minimum reliability. (This method will be discussed in greater detail in Chapter V). The test papers of the 500 students taking Test I in May, 1950, were used for item analysis. These item analysis data were utilized in the construction of Test IA.

In order to determine whether there was a difference in the ability to think scientifically before and after the completion of the course in Biological Science, Test I was administered in September, 1950, to 240 students who had had

no biological science at the college level. These students were beginning their first term of the three-term sequence of biological science. This group was also different from any previously mentioned. Of this group 144 were males and 86 were females. The age range was 17 years to 34 years, with a mean age of 19.18 years. The mean and the standard deviation of the test scores were calculated for this group. The reliabilities of the test were determined by the split-half method and by the Kuder-Richardson formula. As there was no means of predicting the exact length of a test of this nature to fulfill the time requirement of one hour and fifty minutes, the number of items used was purely arbitrary. The actual execution of the test indicated that it was too long for all students to complete. Of the 500 students taking the examination in May, 1950, 54 or 10.8 percent failed to finish in the allotted time. Of the 240 students taking the test in September, 1950, 24 or 10 percent failed to complete the test. Since the test was too long the poorer items, as determined by item-analysis, were eliminated. The remaining items constituted Test IA.

This final form of the test, consisting of 125 items, was administered to 330 students at the beginning of the three-term sequence of biological science in September, 1950. This is a different group from any previously mentioned in this study, and included 182 males and 148 females. The age range was from 16 years to 38 years with a mean of 18.62

years. Thirteen, or 3.7 percent, did not complete the test. The mean and standard deviation of the test was calculated for this group. The reliability of the test was estimated for this group by correlating the scores made on the odd-numbered items with those made on the even-numbered items. This correlation was corrected by the Spearman-Brown formula. The minimum reliability for this group was estimated by means of the Kuder-Richardson formula.

The seventh step in the construction of the test was its validation. The most important characteristic of a test is its validity<sup>6</sup> which may be defined as the extent to which a test measures what it purports to measure.<sup>7</sup> Chapter VI is devoted to a discussion of this characteristic of the test. The curricular validity of the test was based on the following considerations: (1) designing the test to measure the specific behaviors which attend the steps of scientific thinking, (2) submitting the test to qualified judges for criticism, and (3) using free responses of students as foils wherever feasible.

The test was validated statistically by correlating total scores made on the battery of tryout tests with such traits as (1) intelligence, (2) reading ability, and (3)

---

<sup>6</sup> Herbert E. Hawkes, E. F. Lindquist and C. R. Mann, The Construction and Use of Achievement Examinations. Cambridge, Mass.: Houghton Mifflin Company. 1936. p. 21.

<sup>7</sup> Adkins, op. cit., p. 160.

knowledge of biological facts. As previously mentioned, psychological examination scores and reading test scores were available for 264 of the 500 students who took Test I, The Ability to Think Scientifically, in the spring of 1950. These scores were correlated with the scores made by these students on Test I.

Another method of validating the test was the comparison of the scores made by students on Test I at the beginning of the course in Biological Science with the scores made by another group after taking three quarters of Biological Science. The assumptions underlying this comparison will be discussed in Chapter VI. Test IA was administered to 136 students at the beginning and at the end of the first quarter of the three-term Biological Science sequence. The scores made by these students at these two times were compared.

Scores made by a group of 143 students on Test IA were compared with ratings of these students on their ability to think scientifically. The ratings were made by the instructors who taught these students in Biological Science. The rating sheet and the methods used to obtain scores from these ratings and the statistical treatment of these data will be discussed in detail in the chapter on the validation of the test.

## CHAPTER IV

### THE DEVELOPMENT OF THE TEST ITEMS

This chapter is devoted to a discussion of those steps in the construction of the test which preceded and included the writing of the preliminary items which were used in the tryout tests. These steps included the formulation of the educational objectives to be tested, the definition of the behaviors which attend these objectives, the identification of situations in which the students could be expected to display the types of behaviors identified in step two and the writing of items designed to appraise the behaviors identified.

### THE FORMULATION OF THE EDUCATIONAL OBJECTIVES

The overall objective to be measured by the test was the ability to think scientifically. As discussed in Chapter II, scientific thinking involves a number of elements. The major elements as outlined by Keeslar<sup>1</sup> have been reworded and are presented here as the major objectives involved in the ability to think scientifically.

1. The ability to sense a problem.
2. The ability to state a problem.

---

<sup>1</sup> Oreon Keeslar, "The elements of scientific method." Science Education, 29:273-278, December, 1945.



3. The ability to delimit a problem.
4. The ability to recognize facts which are related to the problem.
5. The ability to formulate hypotheses.
6. The ability to plan experiments to test hypotheses.
7. The ability to carry out experiments.
8. The ability to interpret data.
9. The ability to formulate generalizations based on data.
10. The ability to apply generalizations to new situations.

Some of the above abilities are creative, others are critical, while others involve both critical and creative aspects of scientific thinking. For example, the sensing of a problem is a creative activity. So also is the actual formulation of hypotheses, but the detecting of illogical hypotheses is a critical activity. The planning of experiments also has both creative and critical aspects. As Burke<sup>2</sup> points out, there is overlapping between critical and creative thinking, and the decision as to where to draw the line must be based on pragmatic considerations. Thus, he included the drawing of valid inferences from data as critical thinking since it may be measured by objective tests. The behaviors which have been considered primarily critical will

---

<sup>2</sup> Paul J. Burke, "Testing for critical thinking in physics," American Journal of Physics, 17:527-532, December, 1949.

be discussed in detail in a later portion of this chapter. The tests designed in this study have been limited to the appraisal of the critical aspects of scientific thinking because no method for evaluating the creative behaviors was found in the literature, nor did the writer find it possible to devise satisfactory methods for evaluating these creative aspects of thinking.

According to Burke<sup>3</sup> critical thinking is an abstraction and can have concrete meaning only when applied to some subject matter. Therefore, the behaviors which constitute the elements of critical thinking must be thought of in relation to some specific field; in this instance, the field was biology.

#### THE DEFINITION OF THE BEHAVIORS

Methods used to determine the behaviors. In order to determine the kinds of behaviors attending the steps in the scientific method several approaches were used.

The lists of steps in scientific thinking as presented by Keeslar<sup>4</sup> and as presented in the 46th Yearbook,<sup>5</sup> both of which were reviewed in Chapter II, offered a source

---

<sup>3</sup> Burke, loc. cit.

<sup>4</sup> Keeslar, op. cit., pp. 273-278.

<sup>5</sup> Science Education in American Schools. Forty-Sixth Yearbook of the Society for the Study of Education, Part I. pp. 145-147. Chicago: University of Chicago Press, 1947.

for the definition of many of the behaviors involved in scientific thinking. The major steps constituted the primary objectives while the minor steps, in many cases, implied specific behaviors which could be measured.

A second source of behaviors was literature on tests and test construction, committee reports on behaviors involved in scientific thinking, and reports of research on behaviors of persons doing scientific research.

In his book on the construction of achievement tests, Tyler<sup>6</sup> discussed tests to measure the ability to use the scientific method and the ability to infer. In these sections he described some of the behaviors involved. This was a rather early piece of work in the area of definitions of behaviors and was included here more for its historic interest than for its value as a source of behaviors.

Hawkes, Lindquist and Mann<sup>7</sup>, in a chapter on examinations in the natural sciences, discussed some of the behaviors which give evidence of the student's ability to use reliable sources of information, to recognize unsolved problems, to draw reasonable generalization from data, and to plan experiments.

A very useful source of behaviors involved in

---

<sup>6</sup> Ralph W. Tyler, Constructing Achievement Tests. Columbus, Ohio: Ohio State University. 1934. pp. 24-30.

<sup>7</sup> Herbert E. Hawkes, E. F. Lindquist and C. R. Mann, The Construction and Use of Achievement Examinations. Cambridge, Mass.: Houghton Mifflin Company. 1936. pp. 231-247.

scientific thinking was "Science in General Education."<sup>8</sup> A portion of one chapter of this book is devoted to a discussion of the nature of reflective thinking. Another chapter is devoted almost entirely to the evaluation of students growth in reflective thinking. Situations are described which show the kinds of behaviors expected of students who are proficient in the ability to think reflectively. The objectives analyzed are: (1) the ability to discover and define problems, (2) the ability to observe accurately, (3) the ability to select facts relevant to a problem, (4) the ability to collect and organize facts, (5) the ability to draw inferences from facts, (6) the ability to recognize proof, and (7) the ability to plan experiments to test hypotheses.

In the report on the methods of evaluating student progress in the Eight-Year Study, Smith and Tyler<sup>9</sup> discuss in detail the behaviors involved in the students ability to interpret data and in some detail the behaviors involved in an understanding of the nature of proof. The behaviors involved in the ability to interpret data were derived from discussions of the committee on the interpretation of data.

---

<sup>8</sup> Progressive Education Association, Science in General Education. New York: D. Appleton-Century Company. 1938. pp. 393-412.

<sup>9</sup> Eugene R. Smith, Ralph W. Tyler and the Evaluation Staff, Appraising and Recording Student Progress. New York: Harper and Brothers. 1942. pp. 38-41, 126-130.

The committee was comprised of a representative from each school interested in this objective, and the members of the Evaluation Staff of the Eight-Year Study. The work of this committee was quite exhaustive. Most of the behaviors listed under interpretation of data in the list of behaviors presented in this thesis are either mentioned or implied in Smith and Tyler's discussion of behaviors involved in their discussion of the interpretation of data and their discussion on the nature of proof.

Johnson,<sup>10</sup> in a discussion of her test of straight thinking, presents the kinds of behaviors which her test purported to measure. The major abilities discussed are: (1) the ability to analyze a problem, (2) the ability to interpret data, (3) the ability to evaluate arguments, (4) the ability to test hypotheses through reasoning, and (5) the ability to recognize valid causal relationships.

The Committee on Research in Secondary School Science<sup>11</sup> focused its attention on the development of problem-solving as the area in which research was needed. The members of this committee considered problem-solving to be a general type of human behavior which included specific, inter-

---

<sup>10</sup> Alma Johnson, "An experimental study in analysis and measurement of reflective thinking." Speech Monographs, 10:83-96, (Annual) 1943.

<sup>11</sup> Committee on Research in Secondary School Science, "Problem-solving as an objective of science teaching." Science Education, 33:192-195, April, 1949.

related behaviors. They analyzed these behaviors in the following areas::

1. Behaviors concerned with the identification of problems.
2. Behaviors related to the establishment of facts about the problem.
3. Behaviors related to the formulation of hypotheses.
4. Behaviors related to the testing of hypotheses.
5. Behaviors concerned with the results of testing hypotheses.

The behaviors listed by this committee were incorporated into the list of behaviors presented in the present study.

Burke,<sup>12</sup> in discussing the development of test items to test the ability to think scientifically, says that before any test of critical thinking could be constructed, or before any orderly attempt could be made to teach the scientific method, the concept must be made more precise than it has been previously. He presents an operational definition consisting of a set of about 30 behaviors. He offers the list as a tentative definition. Most of the behaviors in his list have been incorporated in the outline of behaviors presented in this chapter.

A study sponsored by the American Institution of

---

<sup>12</sup> Burke, op. cit., pp. 27-32.

Research and the American Council on Education and supervised by Flanagan,<sup>13</sup> was made of the activity of research workers on the job, to identify and define the characteristics of effective scientific personnel, in terms of specific observations and records of the work behavior of these personnel.

The method used to obtain these behaviors was not the opinions or beliefs of supervisors of research, but rather the actual experiences, in the form of reports of behavior which led to success or failure of individuals on various parts of their jobs. Reports of what actually happened were turned in to the committee. About 500 research workers were contacted, who were asked to describe critical incidents in which a person had been effective or ineffective in research techniques. Upon the completion of the interviews the behaviors described were classified into groups of similar behaviors.

On the basis of the classification of the behaviors a comprehensive check list was prepared for the evaluation of research workers. Each area was divided into sub-areas. In addition to the check list which included descriptions of effective and ineffective behavior in each of the areas, definitions of the areas were written to provide a general description of the content of the area.

---

<sup>13</sup> John C. Flanagan, Critical Requirements for Research Personnel. Pittsburg: American Institute for Research. 1949. pp. 24-39.

Area I was the formulation of hypotheses and problems. This area was defined as stressing creative behavior, and included the sensing and exploring of new problem areas, delimiting problems and the proposing of hypotheses to fit the available facts. Within this area 21 effective and eleven ineffective types of behaviors were described. These made up the items of the check list.

Area II dealt with the planning and designing of an investigation; Area III was concerned with the conducting of the investigation and Area IV was the interpretation of research results. Areas V, VI, VII, and VIII were not related to scientific thinking but dealt with preparing reports, administration of research, organizational responsibility and personal responsibility and were not related to the present investigation.

Although this work was outstanding in its thoroughness and although over 100 behaviors relating to research ability were presented, most of them have not been incorporated into the outline presented in this chapter because many were creative activities, and many others were manipulative activities. The critical activities, however, were incorporated into the outline of behaviors which will be presented later in this chapter.

A third source used in the identification of behaviors involved in scientific thinking was the interviewing



of some of the members of the department of Biological Science at Michigan State College. These persons were asked to describe the behaviors they had observed in students whom they believed to show considerable ability to think scientifically, and the kinds of behaviors they had observed in students who seemed to them to be very inferior in their ability to think scientifically. The major abilities mentioned in these interviews were the ability to devise and evaluate experiments, and the ability to interpret data. Specific behaviors were described. (These will be discussed in greater detail in Chapter VI where a description of the ratings sheet devised for the validation of the test will be discussed.)

The final source used in the definition of behaviors was the experience of the writer as an instructor in the course of Biological Science at Michigan State College and her experience as a member of the committee responsible for the construction of departmental examinations.

An Outline of the Behaviors. Below is an analysis, in outline form, of the types of behaviors involved in scientific thinking which it was believed could be measured by objective tests. It is not assumed that this is an all-inclusive list. It is, however, a synthesis of the behaviors identified from the above mentioned sources.

## 1.00 Ability to recognize problems.

- 1.10 Ability to recognize a problem or a perplexity in the context of a paragraph or an article.
- 1.20 Ability to distinguish between a fact (observation) and a perplexity or problem.
- 1.30 Ability to recognize a problem even when it is stated in expository form rather than in interrogatory form.
- 1.40 Ability to distinguish a problem from a possible solution to a problem (hypothesis) even when the hypothesis is presented in interrogatory form.
- 1.50 Ability to avoid becoming diverted from the major problem into side issues.

## 2.00 Ability to delimit a problem.

- 2.10 Ability to distinguish between major and minor problems.
- 2.20 Ability to isolate the single major problem or single major idea in a problem.
- 2.30 Ability to see the relationship of minor problems to the major problems.
- 2.40 Ability to distinguish between relevant and irrelevant problems.
- 2.50 Ability to analyze the problem into its essential parts.
- 2.60 Ability to concentrate on the main problem.
- 2.70 Ability to recognize the basic assumptions of a problem.

## 3.00 Ability to recognize and accumulate facts related to the solution of a problem.

- 3.10 Ability to select the kind of information needed to solve the problem.
- 3.20 Ability to recognize valid evidence.
- 3.30 Ability to differentiate between reliable and unreliable sources of information.
- 3.40 Ability to select data pertinent to the solution of the problem.
- 3.50 Ability to recognize the difference between data pertinent to the solution of the problem and that which is unrelated.

## 4.00 Ability to recognize an hypothesis.

- 4.10 Ability to distinguish an hypothesis from a problem.
- 4.20 Ability to differentiate between a statement that describes an observation and a statement which is an hypothesis about the fact.
- 4.30 Ability to distinguish between an hypothesis as a possible solution to a problem and a conclusion (probable solution to a problem).
- 4.40 Ability to recognize the tentativeness of an hypothesis.

5.00 Ability to plan experiments to test hypotheses.

- 5.10 Ability to select the most reasonable hypothesis to test.
- 5.20 Ability to differentiate between an uncontrolled observation and an experiment involving controls.
- 5.30 Ability to recognize the fact that only one factor in an experiment should be variable.
  - 5.31 Ability to recognize what factors must be controlled.
  - 5.32 Ability to recognize the overall control.
  - 5.33 Ability to recognize the partial controls.
  - 5.34 Ability to recognize the variable factor.
  - 5.35 Ability to understand why the overall control was included in an experiment.
  - 5.36 Ability to recognize the factor being held constant in the overall control.
  - 5.37 Ability to recognize the factors being held constant in the partial controls.
- 5.40 Ability to recognize experimental and technical problems inherent in the experiment.
- 5.50 Ability to criticize faulty experiments when:
  - 5.51 The experimental design was such that it could not yield an answer to the problem.
  - 5.52 The experiment was not designed to test the specific hypothesis stated.
  - 5.53 The method of collecting the data was unreliable.
  - 5.54 The data were not accurate.
  - 5.55 The data were insufficient in number.
  - 5.56 Proper controls were not included.
  - 5.57 No controls were included.

6.00 Ability to carry out experiments.

- 6.10 Ability to recognize existence of errors in measurement.

- 6.20 Ability to recognize when the precision of measurement given is warranted by the nature of the problem.
- 6.30 Ability to make accurate observations.
  - 6.31 Ability to observe differences in situations which are similar.
  - 6.32 Ability to observe similarities in situations which are different.
- 6.40 Ability to organize facts into table, graphs, etc. for easy interpretation.
- 7.00 Ability to interpret data.
  - 7.10 Ability to handle certain basic skills necessary to the interpretation of data.
    - 7.11 Ability to read tables and graphs.
    - 7.12 Ability to perform simple computations.
  - 7.20 Ability to evaluate relevancy of data.
    - 7.21 Ability to recognize hypothesis and conclusions contradicted by the data.
    - 7.22 Ability to recognize hypotheses and conclusions which are unrelated to the data.
    - 7.23 Ability to select the hypothesis from a group of hypotheses which most adequately explains the data.
    - 7.24 Ability to recognize facts which support an hypothesis or a conclusion.
    - 7.25 Ability to recognize facts which contradict an hypothesis or a conclusion.
  - 7.30 Ability to differentiate between facts and inferences.
    - 7.31 Ability to differentiate between an observation and a conclusion drawn from the observation.
    - 7.32 Ability to differentiate a conclusion from an hypothesis.
    - 7.33 Ability to distinguish an assumption upon which a conclusion depends and the conclusion itself.
    - 7.34 Ability to distinguish a fact from an assumption.
  - 7.40 Ability to recognize the limitations of data.
    - 7.41 Ability to differentiate between what is established by the data alone and what is implied by the data.
    - 7.42 Ability to recognize that a statement which goes beyond the data cannot be absolutely true.
    - 7.43 Ability to recognize that generalizations from results of an experiment can only be extended to new situations when there is considerable similarity between the situations.
    - 7.44 Ability to confine definite conclusions to the evidence at hand.

- 7.50 Ability to consider as possibly true or probably true inferences based on the data.
- 7.51 Ability to make inference on the basis of trends.
- 7.52 Ability to extrapolate.
- 7.53 Ability to interpolate.
- 7.54 Ability not to be so overcautious that all statements which go beyond the data are rejected because of insufficient evidence.
- 7.60 Ability to perceive relationships in data.
- 7.61 Ability to make comparisons.
- 7.62 Ability to see element in common to several items of data.
- 7.63 Ability to recognize prevailing tendencies and trends in data.
- 7.64 Ability to recognize that when two things vary together that there may be a relationship between them, but does not assign cause and effect judgments on the basis of this relationship.
- 7.65 Ability to formulate reasonable generalizations based upon the data.
- 7.70 Ability to recognize the nature of evidence.
- 7.71 Ability to recognize the difference between direct and indirect evidence.
- 7.72 Ability to recognize a statement which is given as evidence as not being evidence when the statement contradicts the conclusion.
- 7.73 Ability to recognize a statement which is given as evidences as not being evidence when the statement is unrelated to the conclusion.
- 7.74 Ability to recognize evidence for an inference and to choose such evidence from a series of statements.
- 7.75 Ability to recognize the validity of the evidence used to support conclusions.
- 7.80 Ability to recognize the assumptions involved in the formulation of hypotheses and conclusions.
- 7.81 Ability to recognize assumptions which go beyond the data but which are essential to the formulation of an hypothesis.
- 7.82 Ability to recognize assumptions which must be maintained in the drawing of a conclusion.
- 7.83 Ability to recognize assumptions which can be checked experimentally.
- 7.84 Ability to recognize invalid assumptions.

8.00 Ability to apply generalizations to new situations.

- 8.10 Ability to refrain from applying generalizations to new situations when the new situation does not closely parallel the experimental situation.
- 8.20 Ability to be aware of the tentativeness of predictions about new situations even when there is a close parallel between the two situations.
- 8.30 Ability to recognize the assumptions which must be made in applying a generalization to a new situation.

#### THE LOCATION OF THE SOURCE MATERIALS FROM WHICH THE ITEMS COULD BE CONSTRUCTED

The third step in the development of the test was the identification of situations in which the student could be expected to display the types of behaviors implied in the steps of scientific thinking. Each major objective was considered and situations were considered which might be utilized in the construction of items to test the abilities involved in these objectives.

There were certain requirements which should be met in the selection of the material. It was considered reasonable that in all cases the material should be (1) of some interest to the student, (2) free from technical terms, (3) comprehensible to the student who had had no training in biology, (4) on biological subjects, and (5) obtained from valid sources.

It was thought that the abilities involved in the recognition of a problem, an hypothesis, a fact, and a conclusion could be discovered by having a student actually locate them in his reading. In the development of an

objective test it seemed that one way in which these behaviors could be measured was by the presentation of short essays or paragraphs which contained problems, etc. and having the student identify them. With this in mind, popular and scientific journals were inspected for descriptions of experiments or observations which contained problems, hypotheses, experiments, observations and conclusions.

These were judged by the following criteria:

1. They should be of such a nature that they could be condensed into a paragraph or two.
2. They should each contain a problem or problems, hypotheses, observations and experiments, and a conclusion.

It was tentatively assumed that a student's ability to delimit a problem might be measured by giving him a comprehensive problem so stated that it could not be solved unless it were broken down into a series of minor problems. Such problems were located in textbooks, research journals, and by interview of members of the Department of Biological Science of Michigan State College. The criteria used in the selection of the problems were:

1. Unsolved problems were chosen so that the student could not know the solution to the problem.
2. The problems should be broad major problems.

In order to measure a student's ability to plan experiments it was necessary to locate problems and hypotheses already under investigation or those which might be investigated, thus limiting the possibility of the student having

had experience with the problem. Some of these were found in research journals and some were obtained by interviews with staff members of the Department of Biological Science at Michigan State College. The criteria by which they were judged were:

1. They should be of such a nature that no technical apparatus would be needed to design an experiment.
2. They should be within the experience of the student; that is, the general problem should deal with situations which could reasonably be assumed to be familiar to him.

In order to test a student's understanding of experimental design actual experiments were located in which the student could identify controls, partial controls, etc. These experiments were located in scientific journals. It was assumed that the experiments should be:

1. Entirely new to the students.
2. On a subject with which the student was familiar.

These assumptions were met by choosing experiments from technical journals which the average student would not have read, and by choosing experiments which were about rather common subjects, such as food, plants, etc.

It was thought that the ability to organize data could be tested by giving students raw data to graph. A search of textbooks and journals produced this type of material. The criteria used to judge the usability of the data were:



1. The data must be in units familiar to the student.
2. The data must be such that only few points would be needed to plot a curve so that a number of curves could be plotted in a minimum of time.

Scientific journals and advanced textbooks were examined for data which the student could interpret. It was assumed that these data should be entirely new to the student.

#### THE CONSTRUCTION OF THE EVALUATION INSTRUMENTS

The fourth step in the development of the test was the selection of promising techniques, and the inventing of new techniques to obtain evidence concerning the attainment of the objectives. Previous tests designed to test certain phases of scientific thinking were examined. No tests for biology were found which measured all of the objectives listed. There were only a few which measured any of the objectives. New techniques for appraising the desired behaviors were devised, paragraphs from sources were rewritten, students were presented with some of the materials identified in step three for free responses which were culled and classified. On the basis of this work nine tryout tests were devised. The following discussion gives in more detail the method used in the construction of each of the instruments and the objectives and types of behavior which each was intended to evaluate.

In the development of the test items certain requirements regarding mechanics were set up. The first requirement was that the test be easily scored. A five-response machine scored answer sheet was chosen as the most appropriate for the purposes of this test. A second consideration was the test form. A five-choice key was selected as the most suitable form inasmuch as a single key for each test would enable the student to answer a rather large number of items in a fairly short time, thus increasing the reliability of the test. He would become acquainted with the key and thus reduce the reading time of the test. Each tryout test had a separate key.

After the test items had been constructed they were given to five experts for keying, criticism and suggestion. The items were revised on the basis of these judgments, and assembled into tryout tests. (See Appendix I.)

The first tryout test, hereafter referred to as Test A, was designed to evaluate the student's ability to recognize problems, hypotheses, experimental conditions and conclusions. Five paragraphs were written, each on a different subject and each based on short articles from popular magazines. Certain parts of the paragraphs were underlined; these underlined portions, preceded by a number indicating the item number, constituted the 74 items of the test.

The directions given to the student, the key for the test and a portion of one of the paragraphs follows:

## TEST A

## SOME STEPS IN SCIENTIFIC THINKING

This test is designed to measure your ability to differentiate phases of thinking. These steps include major problems or perplexities, possible solutions to problems, observations which are not results of experimentation but rather preliminary observations, results of experimentation, and conclusions.

Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space on the answer sheet.

Key

1. A major problem (either stated or implied).
2. Hypothesis (possible solution to problem).
3. Results of experimentation.
4. Observations (not experimental).
5. Conclusion (probable solution to problem).

Ever since the days of Hippocrates one of medicine's big mysteries has been (1) the bodily process that transforms disease into death. With a special type of equipment which makes blood vessels transparent and three dimensional under a microscope, one investigator began examining the blood of healthy animals. The (2) blood cells of the healthy animals are separate and move rapidly. One day while observing the blood of a monkey dying of malaria, this researcher saw that the (3) blood was flowing slowly.

Test B, designed to test the student's ability to delimit problems, was constructed from free responses of students. For example, several facts about colds were given to the students. They were asked to read the paragraph and

state briefly the problem or problems presented. The major problem was: What causes colds? In constructing the test this problem was followed by other problems which the students had suggested. Four major problems were presented; each of which was followed by a series of questions. There was a total of 67 such questions in this tryout test. A portion of Test B follows:

## TEST B

### THE DELIMITATION OF PROBLEMS

This portion of the test is designed to test your ability to delimit a problem. A problem is presented. This is followed by a series of questions. Rate the questions according to the following key.

#### Key

1. This question must be answered in order to solve the problem.
2. This question if answered might be useful in the solution of the problem.
3. The answer to this question, though related to the problem, would not help in the solution to the problem.
4. This question is completely unrelated to the problem.
5. This question if answered in the affirmative is a basic assumption of the problem.

PROBLEM: What causes colds?

QUESTIONS:

1. Do all people have colds?

Test C was designed to measure the student's understanding of the experimental method. This test was also constructed on the basis of free responses from students. They were presented with a problem and hypotheses and were

instructed to design an experiment to test each hypothesis presented. For example: Problem: What are some of the requirements of sprouting seeds? Hypothesis: Oxygen is a requirement of sprouting seeds.

The papers were cut so that the experiments designed to test a single hypothesis could be sorted and these were placed in piles according to the key which was used in Test C. Some of the responses were satisfactory experiments, others were faulty for one reason or another, some were faulty for several reasons. Those which were faulty in more than one way were discarded. Ten or eleven responses for each problem were chosen as the test items. Six series of experiments with a total of 62 items constituted Test C, a portion of which is presented here:

## TEST C

### EXPERIMENTAL PROCEDURES

This test is designed to measure your ability to recognize faulty experimental procedures and to test your ability to select the best of a series of experiments. In each case a problem and a possible solution to the problem (an hypothesis) are presented. In each case the experiments were designed by students to test the hypotheses. Judge each experiment according to the following key.

#### Key

1. This experiment is satisfactory.
2. This experiment is unsatisfactory because it lacks a control or comparison.
3. This experiment is unsatisfactory because the control or comparison is faulty.
4. This experiment is unsatisfactory because it is unrelated to the hypothesis.

5. None of the above - the experiment or situation is unsatisfactory for reasons other than those listed in 2, 3, and 4.

PROBLEM: What are some of the requirements for the sprouting of seeds?

HYPOTHESIS: Oxygen is a requirement for the sprouting of seeds.

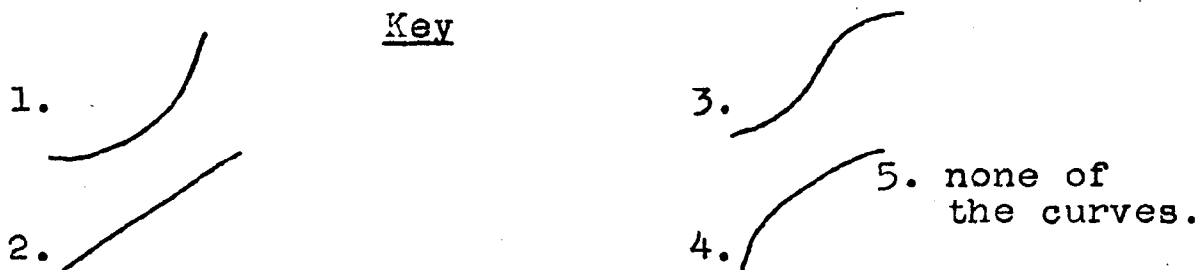
1. Plant one seed in a container where oxygen is available and place another seed in a container where all oxygen has been removed. Keep all other conditions the same.

Test D, designed to measure the student's ability to organize data, contained twenty items similar to the one illustrated here:

### TEST D

#### ORGANIZATION OF DATA

This test is designed to test your ability to organize data. Select from the key below the curve which best fits the data. If none of the curves fit the data mark space five on your answer sheet.



1. The horizontal axis represents temperature. The vertical axis represents the amount of Substance A derived from Substance B.

Temperature

Amount of Substance A

10°C.  
25°C.  
35°C.  
60°C.

4 grams  
7 grams  
9 grams  
14 grams

Test E is similar to one described by Engelhart and Lewis.<sup>14</sup> It was designed to measure the student's understanding of the relation of facts to the solution of a problem. All of the 74 items of this test were related to the overall problem: What factors are involved in the transmission and development of Infantile Paralysis (Poliomyelitis)? Six hypotheses were presented. Each hypothesis was followed by a series of facts which constituted the items. The data for the test were obtained from articles on infantile paralysis in research journals and medical journals. A portion of Test E follows:

### TEST E

#### EVALUATION OF HYPOTHESES

This test is designed to measure your understanding of the relation of facts to the solution of a problem. The overall problem involved in this test is presented. This is followed by a series of possible solutions to the problem (hypotheses). After each hypothesis there are a number of items, all of which are true statements of fact. Determine how the statement is related to the hypothesis and mark each statement according to the key which follows the hypothesis.

GENERAL PROBLEM: What factors are involved in the transmission and development of Infantile Paralysis (Poliomyelitis)?

HYPOTHESIS I: In man the disease is contracted by direct contact with persons having the disease.

---

<sup>14</sup> Max D. Engelhart and Hugh B. Lewis, "An attempt to measure scientific thinking." Educational and Psychological Measurement, 1:289-294, Third quarter, 1941.

Key

For items 1 through 11 mark space if the item offers:

1. Direct evidence in support of the hypothesis.
2. Indirect evidence in support of the hypothesis.
3. Evidence which has no bearing on the hypothesis.
4. Indirect evidence against the hypothesis.
5. Direct evidence against the hypothesis.

1. Monkeys free from the disease almost never catch infantile paralysis from infected monkeys.
2. Most strains of infantile paralysis virus can be transferred from man only to monkeys and apes and not to other animals.
12. What is the status of hypothesis I?
  1. It is true.
  2. It is probably true.
  3. It is false.
  4. It is probably false.
  5. The data are contradictory, hence its truth or falsity cannot be judged.

Test F was designed to measure the student's ability to interpret data and to test his understanding of experimentation. The directions for this tryout test and a portion of the test are given below:

## TEST F

## EXPERIMENTATION AND THE INTERPRETATION OF DATA

This test was designed to measure your ability to interpret data and to test your understanding of experimentation. In each case the numbers in the first column are the numbers which you will use as your answer. Thus the table presented becomes both the source of data and your key for the questions which follow it. In each case where a test tube number or group number is called for the one which gives positive evidence for the statement should be given. Below this the control or comparison is called for. This is the test tube or group number of the data which offers a comparison. For example:



1. Leaf in dark - no starch.
2. Leaf in light - starch.

Light is necessary for the production of starch. You would mark space 2 because this is the positive evidence, but it would be meaningless if it were not compared with the leaf in the dark. Therefore, the following item, "What is the control (comparison) for item 1?", would be marked space 1.

Items 1 through 15 refer to the data presented below. Some test tubes were set up and each contained 1 gram of fat. They were marked 1, 2, 3, 4, and 5. Mark each item according to the test tube number called for. Various substances were added to the tubes containing fat. All substances were dissolved in water before they were added to the fat. All test tubes were kept at 85° F. (Water boils at 212° F.) For test tube 5, Substance A was boiled and then allowed to cool before it was added to the fat.

Test Tube Number	Content of tube	Amt. of Substance B present after 24 hours
1	Fat plus Substance A	.1 gram
2	Fat plus Substance A plus Substance C	.5 gram
3	Fat plus Water	.0 gram
4	Fat plus Substance C	.0 gram
5	Fat plus Substance A (boiled)	.0 gram

1. Give the number of the test tube which acts as a control (comparison) for the entire experiment.
2. Give the number of the tube which gives evidence that fat does not break down spontaneously into Substance B in 24 hours.
3. Give the number of the tube used to show that a temperature of 85° F. was not sufficient to cause fat to be broken down into Substance B.
4. Give the test tube number of the tube which gives evidence that Substance A is the active substance in the breakdown of fat to Substance B.
5. Give the test tube number of the tube which is the control (comparison) for item # 4.

Five such series of items were included in Test F. The total number of items was 72.

Test G is somewhat like the test described by Teichman<sup>15</sup> which was constructed to evaluate conclusions in terms of reasonableness, sufficiency and pertinent data. This test was constructed from free responses of students. A problem was presented. This was followed by data. For example: A student was interested in developing a test for a certain substance. In all 100 cases his test was positive. The students were requested to state a conclusion. In some instances, as in the above, there was no control included so no conclusion was really possible. Some of the students realized this; others wrote conclusions. The answers were sorted into stacks according to the key used for Test G. The most appropriate responses were chosen as the 100 items for the test.

### TEST G

#### DRAWING OF CONCLUSIONS

This test was designed to measure your ability to make conclusions. When facts are analyzed and studied they sometimes yield evidence which help in the solution of a problem. However, any conclusion must be checked before it can be accepted. The following key includes four ways in which conclusions may be faulty. Each of

---

<sup>15</sup> Louis Teichman, "The ability of science students to make conclusions." Science Education, 28: 268-279, December, 1944.

the items present a question or problem, a brief description of an experiment and one or more conclusions drawn from the experiment. Each experiment was repeated many times. Read each problem, experiment and the conclusions. Where several conclusions are given evaluate each conclusion separately. Is the conclusion tentatively justified by the data? If so, mark space 1 on your answer sheet. If the conclusion is not justified determine whether 2, 3, 4, or 5 in the key is the best reason for it being faulty and mark the proper space on your answer sheet.

### Key

The conclusion is:

1. Tentatively justified.
2. Unjustified because it does not answer the problem.
3. Unjustified because the experiment lacks a control (comparison).
4. Unjustified because the data are faulty or inadequate, though a control was included.
5. Unjustified because it is contradicted by the data.

PROBLEM: A student was interested in developing a test for a certain type of substance. In all 100 cases his test was positive.

1. He concluded that the test was a specific test for the substance.

The final tryout test was in reality two tests, Test H and Test J, combined into one. In all, these tests contained 168 items. Test H was devised to measure the student's ability to interpret data. Data were presented to the students. These were followed by a series of items which were possible interpretations, restatements, explanations, extensions, and comparisons of the data. These items constituted Test J.

## TEST H

## INTERPRETATION OF DATA

This test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

In freezing of vegetables the common practice for both commercial and home frozen vegetables is to scald the vegetables first by placing them in boiling water for two to three minutes. The following data were obtained in an experiment which measured the amounts of Vitamin C in fresh vegetables, scalded vegetables before freezing, and vegetables frozen for six months. One group of the frozen vegetables was frozen without first scalding, the other group was first scalded. The Vitamin C content of the frozen vegetables was determined before and after they were cooked. All figures indicate the amount of Vitamin C in mg. per 100 cc.

---

<u>Vegetable</u>	<u>Frozen</u>					
	<u>Fresh</u>	<u>Scalded</u>	<u>Unscalded</u>		<u>Scalded</u>	
			<u>Raw</u>	<u>Cooked</u>	<u>Raw</u>	<u>Cooked</u>
Chard (greens)	60	37	20	2	24	14
Spinach	82	43	10	1	27	16
Peas	29	21	14	10	20	16
Green beans	34	29	25	13	23	17
Lima beans	33	20	26	18	20	14

---

1. Scalding of all vegetables causes destruction of some of the Vitamin C content of the vegetables.
2. Spinach is a good source of Vitamin C.

### TEST J

#### GENERALIZATIONS AND ASSUMPTIONS

Items 16 through 21 are a re-evaluation of some of the items 1 through 15. Re-read items 1, 3, 9, 11, 13 and 15 and determine whether they are generalizations, extensions of data, explanations of the data or merely restatements of the data, etc. Answer each according to the following key:

#### Key

1. A generalization; that is the data says it is true for this situation, a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of cause and effect..
4. A restatement of results.
5. None of the above.

16. Item 1.

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusions are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 1 through 15.

#### Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion 1: The breakdown of Vitamin C proceeds spontaneously but is a relatively slow process at low temperature.

22. Vitamin C is a stable substance.
23. There is order in the universe.

## ANALYSIS OF THE TRYOUT TESTS IN TERMS OF THE BEHAVIORS INVOLVED

Table I has been prepared to indicate which of the behaviors outlined earlier in this chapter each of the try-out tests was designed to measure. The major objectives are presented in the table. These are followed by the behaviors reworded into shorter statements. The tests have previously been described, but the descriptive titles are presented here to facilitate the reading of the table.

Test A	Some Steps in Scientific Thinking
Test B	The Delimitation of Problems
Test C	Experimental Procedures
Test D	Organization of Data
Test E	Evaluation of Hypotheses
Test F	Experimentation and the Interpretation of Data
Test G	Drawing of Conclusions
Test H	Interpretation of Data
Test J	Generalizations and Assumptions

An inspection of this table indicates that an attempt was made to cover most of the behaviors observable in persons employing the critical aspects of scientific thinking in the preliminary test battery. It will be seen that a few of the behaviors were not well covered by the tests, such as the ability to recognize valid and invalid data, valid and invalid sources of data, and the ability to carry out experiments. These were omitted chiefly because little attempt has been made to teach these two objectives in the course in Biological Science at Michigan State College.

TABLE I  
BEHAVIORS MEASURED BY THE TRYOUT TESTS

Behaviors	Tests									
	A	B	C	D	E	F	G	H	J	
Recognizes Problems	X									
1.10 Recognizes problems in context	X									
1.20 Distinguishes fact from problem	X									
1.30 Recognizes problem in expository form	X									
1.40 Distinguishes problem from hypothesis	X									
1.50 Distinguishes problem from side issues		X						X		
Delimits Problem		X								
2.10 Distinguishes major problem from minor ones	X	X								
2.20 Isolates major problem or major idea		X								
2.30 Sees relation of minor problems to major one		X								
2.40 Distinguishes relevant from irrelevant problems		X								
2.50 Analyses problem into essential parts		X								
2.60 Concentrates on main problem		X						X		
2.70 Recognizes basic assumptions of problem		X								
Recognizes Facts Related to solution of problem										
3.10 Selects information needed to solve problem		X			X	X				
3.20 Recognizes valid evidence										
3.30 Recognizes reliable sources of information										
3.40 Selects data pertinent to solution of problem					X	X				
3.50 Distinguishes between pertinent and unrelated data		X			X	X				

TABLE I (continued)

Behaviors	Tests									
	A	B	C	D	E	F	G	H	J	
Recognizes hypotheses	X									
4.10 Distinguishes hypothesis from problem	X									
4.20 Differentiates observation from hypothesis	X							X	X	
4.30 Distinguishes hypothesis from conclusion	X							X	X	
4.40 Recognizes tentativeness of hypothesis	X							X	X	
Plans Experiments			X			X	X			
5.10 Selects proper hypothesis to test										
5.20 Differentiates observation from experiment	X		X							
5.30 Uses single variable factor			X			X	X			
5.31 Controls proper factors						X				
5.32 Recognizes overall control						X				
5.33 Recognizes partial control						X				
5.34 Recognizes variable factor						X				
5.35 Understands reason for overall control						X				
5.36 Recognizes constant factor of overall control						X				
5.37 Recognizes constant factor of partial control						X				
5.40 Recognizes problems inherent in experiment			X				X			
5.50 Criticizes faulty experiments when			X				X			
5.51 Not designed to answer problem			X				X			
5.52 Not designed to test hypothesis			X				X			
5.53 Methods were not reliable										
5.54 Data were not accurate										
5.55 Data were insufficient in number			X				X			
5.56 Proper controls were not included			X				X			
5.57 No controls were included			X				X			



TABLE I (continued)

Behaviors	Tests									
	A	B	C	D	E	F	G	H	J	
Carries out experiments										
6.10 Recognizes measurement errors										
6.20 Recognizes precision of measurement necessary										
6.30 Makes accurate observations										
6.31 Observes differences in similar situations						X				
6.32 Observes similarities in different situations						X				
6.40 Organizes facts for interpretation				X						
Interprets data					X	X	X	X	X	
7.10 Handles skills necessary to interpretation						X				
7.11 Can read tables and graphs						X				
7.12 Can perform simple computations				X		X				
7.20 Evaluates relevancy of data					X	X	X	X		
7.21 Recognizes inferences contradicted by data					X		X	X		
7.22 Recognizes inferences unrelated to data					X		X	X		
7.23 Selects best hypothesis to explain data								X		
7.24 Recognizes facts supporting inference					X	X		X		
7.25 Recognizes facts contradicting inference					X	X		X		
7.30 Distinguishes facts from inferences	X							X	X	
7.31 Distinguishes observation from conclusion	X						X	X	X	
7.32 Distinguishes hypothesis from conclusion	X									
7.33 Distinguishes assumption from conclusion	X									
7.34 Distinguishes fact from assumption	X								X	

TABLE I (continued)

Behaviors	Tests									
	A	B	C	D	E	F	G	H	J	
7.40 Recognizes limitations of data					X			X	X	
7.41 Distinguishes data from what is implied by data					X			X	X	
7.42 Recognizes inferences as not absolutely true					X			X	X	
7.43 Recognizes limitations in applying generalizations								X		
7.44 Confines definite conclusions to evidence					X			X		
7.50 Makes inferences based on data					X			X	X	
7.51 Makes inferences based on trends					X	X		X		
7.52 Makes inferences based on extrapolations						X		X	X	
7.53 Makes inferences based on interpolations						X		X		
7.54 Is not too over-cautious					X	X		X	X	
7.60 Perceives relationships in data					X	X	X	X	X	
7.61 Makes comparisons in data					X	X	X	X	X	
7.62 Sees common elements in data					X	X	X	X	X	
7.63 Recognizes tendencies and trends						X	X			
7.64 Suspends cause and effect judgments					X		X	X	X	
7.70 Recognizes nature of evidence					X					
7.71 Distinguishes direct from indirect evidence					X					
7.72 Recognizes evidence which contradicts conclusion					X	X	X	X		
7.73 Recognizes evidence unrelated to conclusion					X	X	X	X		
7.74 Recognizes evidence for inferences					X	X		X		
7.75 Recognizes validity of evidence					X		X	X	X	

[illegible]

## CHAPTER V

### THE STATISTICAL ANALYSES OF THE TESTS AND THE TEST ITEMS

This chapter is devoted to a presentation of the statistical analyses of the tests and the test items. The means, standard deviations, and reliabilities of each of the tryout tests are presented. Item analysis data for the items in the tryout tests have been summarized. Inter-correlation of the tryout test scores have been calculated and data concerning the degree of overlapping of the tryout tests are discussed. This discussion is followed by analyses of Tests I and IA and by the item analysis data on these tests.

#### METHODS USED IN ITEM-ANALYSIS

Item validity may be defined as a measure of the item's correlation with a criterion.<sup>1</sup> In this case the criterion used was the scores on the tryout test which included the particular item for which the validity was to be determined. The purpose of determining item validity is to identify good items to be retained and poor items to be eliminated or revised. Poor items are generally defined as those lacking in discriminative power while good items are discriminatory

---

<sup>1</sup> Dorothy C. Adkins, Construction and Analysis of Achievement Tests. Washington: U. S. Government Printing Office. 1947. p. 180.

Good items are those missed more often by those persons who have a low degree of the quality being measured, (in this case, the ability to think scientifically), and answered correctly more often by persons having much of this same quality, whereas poor items are answered correctly by the same number of persons, irrespective of their ability. Item validity may be estimated by any one of several methods. Test items are usually validated by comparing the proportion of persons having high scores on the test who answer the item correctly with the proportion of persons having low scores on the test who answer the item correctly. Kelley<sup>2</sup> has shown that the best estimates of the correlation of the item with the total test score can be obtained by using the responses of the upper 27 percent or total score and the lower 27 percent on total score of the group for the calculations.

The estimated item correlation was determined by two methods, both of which required the determination of the percent in the upper 27 percent of the group and the percent in the lower 27 percent of the group answering the items correctly. One method was devised by Flanagan.<sup>3</sup> By this method

---

<sup>2</sup> Truman L. Kelly, "The selection of upper and lower groups for the validation of test items." Journal of Educational Psychology, 30:17-24, January, 1939.

<sup>3</sup> John C. Flanagan, "General Considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution." Journal of Educational Psychology, 30:674-680, December, 1939.

validity is read from a chart, the chart being entered by the percent of successes of each of the groups. The second method used for the estimates of the discrimination power of the items was that of Davis.<sup>4</sup> This method also involves the use of the upper and lower 27 percent of the group. A table is entered by percent of successes of each group; however, the percent successes are calculated differently by Davis' than by Flanagan's method. Straight percent successes are used in the Flanagan method whereas the method devised by Davis involves a correction for guessing. In addition, Davis' method yields a figure which he calls the discrimination index, which is a linear function of the hyperbolic arc tangent of the product-moment coefficient of correlation. He believes that this figure is truly comparable from item to item whereas the coefficient itself is not. The coefficient of correlation cannot justifiably be averaged. However, a table is included in his monograph whereby the discrimination index can be converted to a coefficient of correlation for comparison with results obtained by other methods.<sup>5</sup>

Item difficulties, stated in terms of the percent of persons answering the items correctly, were estimated by the

---

<sup>4</sup> Frederick B. Davis, Item-Analysis Data. Cambridge: Graduate School of Education, Harvard University. 1946. pp. 8-15.

<sup>5</sup> Ibid., pp. 14-15.

method proposed by Davis.<sup>6</sup> These were estimated from the percents of successes of the upper and lower 27 percents of the group. Davis suggests the use of a difficulty index, which like the discrimination index, is read from the table included in his monograph. Like his discrimination index, the difficulty index is corrected for chance. Because item difficulties, when expressed as percents passing the item, cannot justifiably be averaged he devised a difficulty index which is a linear scale. The actual percents can be obtained by use of a table to convert the difficulty indices to percents passing the item.

#### ANALYSES OF TRYOUT TESTS

The tryout tests were administered to 168 students in the spring term of 1950. The tests were scored on the basis of total number of items answered correctly. No correction was made for guessing since students were instructed to answer all items.

##### Analysis of Test A - Some Steps in Scientific Thinking.

Test A, designed to measure an understanding of some of the steps of scientific thinking as described in Chapter IV (see page 126), was comprised of a total of 74 items. The scores on this test ranged from 24 to 67. The mean and its standard error were  $50.60 \pm 0.62$  and the standard deviation and its standard error,  $8.13 \pm 0.44$ . The reliability as estimated

---

<sup>6</sup> Ibid., pp. 2-3.

by the split-half method and adjusted by the Spearman-Brown prophecy formula, was  $.87 \pm .02$ .

Complete item-analysis data for Test A are presented in Table XXXVI of Appendix I. The range of item discrimination, as expressed in terms of estimated coefficients of correlation with total test score, was from .00 to .77. The range in terms of Davis' discrimination indices was from 0 to 61. As previously mentioned, the coefficients of correlation cannot justifiably be averaged, whereas the discrimination indices can be averaged. The mean discrimination index was 29.45. Davis' Table of conversion of indices to equivalent values of coefficients of correlation gave an estimated mean correlation of .45.

The range of difficulty of the items of Test A was from 0 to 95 percent. The range of indices of difficulty was from 0 to 85. Since the difficulty index is subject to statistical treatment these were averaged giving a mean of 55.51. This was equivalent to 60 percent of the group answering these items correctly.

The item analysis data, and the mean of the test indicate that the test was rather easy; the item discrimination data gave evidence that as a whole the items discriminated quite adequately between those students having considerable understanding of the steps of scientific thinking and those not having such an understanding. The reliability coefficient of the test indicated that the test measured whatever quality it was measuring quite consistently. The data for this test are summarized in the following table.



TABLE II  
PERTINENT DATA FOR TEST A

Number of items .....	74
Range of scores .....	24 - 67
Mean .....	50.60 $\pm$ 0.62
Standard deviation .....	8.13 $\pm$ 0.44
Reliability coefficient .....	.87 $\pm$ .02
Range of discrimination indices .....	0 - 61
Mean discrimination index .....	29.45
Range of difficulty indices .....	0 - 85
Mean difficulty index .....	55.51

Analysis of Test B - The Delimitation of Problems.

Test B, devised to measure the ability to delimit problems (see page 127, Chapter IV), as presented originally contained 67 items. Preliminary item analysis revealed that 17 of the items were either lacking in discriminatory power or were negatively discriminating. Since negatively discriminating items reduce the reliability of a test it was deemed advisable to eliminate these 17 items, rescore the papers and on this basis recalculate the item difficulties and item discriminatory values. The scores on the fifty items remaining ranged from 12 to 33; the mean was  $22.46 \pm 0.37$ . The standard deviation was  $4.77 \pm 0.26$  while the estimated reliability coefficient was  $.61 \pm .05$ . Complete item analysis data for this test are presented in Table XXXVII of Appendix I. The range of item discrimination when expressed as an estimated coefficient of correlation was from .04 to .83. The range of discrimination indices was from 4 to 72. The mean discrimination index was 27.08, which when converted to an estimated coefficient of correlation was .44.

The range of difficulty expressed in percent of

successes for Test B was from 4 to 88. The range of the indices of difficulty was from 11 to 75, the mean being 39.40. When converted to percent of successes this became 30 percent. The mean of the test and the percent of successes indicated that this test was relatively difficult. The standard deviation and the range of scores also gave evidence that the items were not all functioning to discriminate between those with superior ability to delimit problems and those inferior in this ability. An inspection of the data presented in Table XXXVII (Appendix I) and of the test items (Appendix I) shows that the most discriminating items of the test were those involving the recognition of the basic assumptions upon which the problem itself rested. This point seemed to be of sufficient interest to present these items separately. The following table gives the discrimination and the difficulty indices of the seven items of the test which purported to measure the student's ability to recognize assumptions underlying problems.

TABLE III

ITEM ANALYSIS DATA ON THE SEVEN ITEMS OF TEST B WHICH MEASURED ABILITY TO RECOGNIZE ASSUMPTIONS UNDERLYING PROBLEMS

Item Number	Discrimination Index	Difficulty Index
5	48	45
9	48	42
21	72	44
28	39	46
38	53	44
45	52	34
58	63	40
mean	53.57	42.14

These items were no more difficult than the other items of the test, in fact, they were answered correctly slightly more often than was the average item, but they were much more discriminating. They accounted, to a large part, for the rather high mean discrimination value of the items of the test. The average estimated coefficient of correlation of these items with the total test score was .71 while the mean difficulty of these items when expressed as percent of successes was 35 percent. The pertinent data for Test B are presented in Table IV.

TABLE IV

## PERTINENT DATA FOR TEST B

Number of items .....	50
Range of scores .....	12 - 33
Mean .....	22.46 $\pm$ 0.37
Standard deviation .....	4.77 $\pm$ 0.26
Reliability coefficient .....	.61 $\pm$ .05
Range of discrimination indices .....	4 - 72
Mean discrimination index .....	27.08
Range of difficulty indices .....	11 - 75
Mean difficulty index .....	39.40

Analysis of Test C - Experimental Procedures. Test C, designed to measure an understanding of experimental procedures (see page 128, Chapter IV), was comprised of 62 items. The scores ranged from 15 to 44; the mean of the test was 26.30  $\pm$  0.41, and the standard deviation was 5.31  $\pm$  0.29. The reliability, as estimated by the split-half method and adjusted by means of the Spearman-Brown prophecy formula was .59  $\pm$  .05.

The item analysis data for Test C are presented in

Table XXXVIII of Appendix I. The range of estimated correlations of the items with the total test score was from  $-.17$  to  $.78$ , the range of discrimination indices was from  $-10$  to  $63$ . The mean discrimination index was  $21.52$  which when changed to an estimated coefficient of correlation was  $.34$ . The range of difficulty indices was from  $0$  to  $59$ ; the mean difficulty index was  $34.37$ , or in terms of percent of success,  $23$  percent. This low percent of success and the low mean of the test both testify to the difficulty of this particular test. The large number of non-functioning items, that is; those with low discriminating power and those answered correctly by sufficiently few students to be accounted for on the basis of chance alone, plus the negatively discriminating items, may account for the rather low reliability of Test C. However, there was a sufficiently large number of satisfactory items in the test to warrant the use of some of the items in the construction of Test I, The Ability to Think Scientifically. Table V is concerned with the pertinent data on Test C.

TABLE V

## PERTINENT DATA FOR TEST C

Number of items .....	62
Range of scores .....	15 - 44
Mean .....	$26.30 \pm 0.41$
Standard deviation .....	$5.31 \pm 0.29$
Reliability coefficient .....	$.59 \pm .05$
Range of discrimination indices .....	$-10 - 63$
Mean discrimination index .....	$21.52$
Range of difficulty indices .....	$0 - 59$
Mean difficulty index .....	$34.37$

Analysis of Test D - Organization of Data. Test D, designed to measure ability to organize data (see page 129, Chapter IV), was comprised of 20 items. The scores on this test ranged from one to ten. The mean of the test was  $10.94 \pm 0.32$ , the standard deviation was  $4.12 \pm 0.23$ . The test had a reliability of  $.93 \pm .01$  as determined by the method of split-halves and correction by the Spearman-Brown formula.

The item analysis data for Test D are presented in Table XXXIX of Appendix I. The range of item discriminations, as expressed by an estimated coefficient of correlation with the total test score, was from .14 to .90; the range of discrimination indices was from 22 to 90. The mean discrimination index was 52.60 which has a corresponding value in terms of coefficient of correlation of .70. The range of difficulty indices was from 22 to 55, the mean being 45.90. This value corresponds to 42 percent successes.

The item analysis data and the mean of the test indicate that the test was of average difficulty. The items were unusually discriminating. As previously mentioned, the tryout test scores were used as the criteria for determining item validity. Since a test score is simply the sum of the scores on individual items the correlation between items and test score is related to the inter-correlations

of individual test items. As pointed out by Conrad,<sup>7</sup> high item validity is an indication that the items are highly consistent or homogenous with other items of the test, and if all of the items are discriminating it means that there is internal consistency or homogeneity of the entire test. Such internal consistency results in a high split-half reliability coefficient. That Test D had considerable internal consistency is shown by the high item validity and the high reliability of the test. An inspection of the test itself also gives evidence of its internal consistence, since the items were all very similar. An inspection of Table I in Chapter IV shows that this test was designed to test a very limited range of behaviors. From the standpoint of item analysis data and test reliability, Test D was the most successful of the tryout tests. However, the fact that it tested a very narrow range of abilities limited its usefulness as a measure of the ability to think scientifically, since this ability includes a wide range of abilities as shown by the analysis of behaviors involved in scientific thinking as presented in Chapter IV. Table VI presents a summary of the pertinent data for Test D.

---

<sup>7</sup> Herbert S. Conrad, Characteristics and Use of Item-Analysis Data. American Psychological Association, Psychological Monographs: General and Applied. No. 295. 1948. p. 15.

TABLE VI  
PERTINENT DATA FOR TEST D

Number of items .....	20
Range of scores .....	1 - 19
Mean .....	10.94 $\pm$ 0.32
Standard deviation .....	4.12 $\pm$ 0.23
Reliability coefficient .....	.93 $\pm$ .01
Range of discrimination indices .....	22 - 90
Mean discrimination index .....	52.60
Range of difficulty indices .....	22 - 55
Mean difficulty index .....	45.90

Analysis of Test E - Evaluation of Hypotheses. Test E was designed to measure the ability to evaluate hypotheses (see page 130, Chapter IV) and was comprised of 74 items. The scores on this test ranged from 15 to 53. The mean of the test was  $34.37 \pm 0.49$  and the standard deviation was  $6.38 \pm 0.35$ . The estimated reliability as calculated by the split-half method and adjusted by the Spearman-Brown formula was  $.71 \pm .04$ .

The item analysis data for this test are presented in Table XXXX of Appendix I. The range of item discriminations expressed in estimated coefficients of correlation of the items with the total test score was from .00 to .71; the range of discrimination indices was from 0 to 54. The mean discrimination index was 24.60 which, when expressed in terms of estimated coefficients of correlation, was .38. The range of difficulty indices was from 0 - 77; the mean was 40.57. This gave a value of 32 percent when expressed as percent of successes.

The items were, as a whole, moderately successful as evidenced by the mean discrimination index. However, the test was somewhat difficult as shown by the fact that the mean of the test was less than half of the total possible points and also by the relatively low mean difficulty index. However, this was also true of most of the tryout tests. Table VII presents a summary of the pertinent data for Test E.

TABLE VII

## PERTINENT DATA FOR TEST E

Number of items .....	74
Range of scores .....	15 - 53
Mean .....	34.37 $\pm$ 0.49
Standard deviation .....	6.38 $\pm$ 0.35
Reliability coefficient .....	.71 $\pm$ .04
Range of discrimination indices .....	0 - 54
Mean discrimination index .....	24.6
Range of difficulty indices .....	0 - 77
Mean difficulty index .....	40.57

Analysis of Test F - Experimentation and Interpretation of Data. Test F was designed to measure the ability to recognize experimental controls and the ability to interpret data (see page 131, Chapter IV). The scores on this test ranged from 18 to 62. The total number of items was 72. The mean of Test F was 47.85  $\pm$  0.66; the standard deviation was 6.48  $\pm$  0.46. The estimated reliability was .89  $\pm$  .02.

Item analysis data for Test F are presented in Table XXXXI of Appendix I. The range of coefficients of correlation with total test scores ranged from .00 to .75. The discrimination indices ranged from 0 to 59; the mean was 30.66.



This gave an estimated mean coefficient of correlation of items with total score of .47. The item difficulties ranged from 0 to 100; the difficulty indices also ranged from 0 to 100, the mean was 55.13. This gave a mean item difficulty of 59 percent of successes.

With the exception of Test D, this test was one of the most successful tests of the tryout battery as evidenced by a relatively high reliability, and by the high item validity. The test was somewhat easier than most of the tests of the tryout battery as shown by the mean of the test and the item difficulty. A summary of the pertinent data for Test F is presented in the following table.

TABLE VIII

## PERTINENT DATA FOR TEST F

Number of items .....	72
Range of scores .....	18 - 62
Mean .....	47.85 $\pm$ 0.66
Standard deviation .....	8.48 $\pm$ 0.46
Reliability coefficient .....	.89 $\pm$ .02
Range of discrimination indices .....	0 - 59
Mean discrimination index .....	30.66
Range of difficulty indices .....	0 - 100
Mean difficulty index .....	55.13

Analysis of Test G - Drawing of Conclusions. Test G,

a hundred item test, was designed to measure the ability to recognize logical conclusions (see page 133, Chapter IV). The scores on this test ranged from 6 to 64. The mean was 38.01  $\pm$  .92; the standard deviation was 11.95  $\pm$  0.65. The estimated reliability of Test G was .90  $\pm$  .01.

Item analysis data for this test are presented in

Table XXXXII of Appendix I. Item validities ranged from  $-.07$  to  $.88$ . Discrimination indices ranged from  $-4$  to  $80$ ; the mean discrimination index was  $31.82$ . This figure represents a mean correlation of  $.48$  of the items with the total test score. The item difficulties ranged from  $0$  to  $89$  percent of successes and the difficulty indices was from  $0$  to  $76$ . The mean difficulty index was  $32.54$  or an average of  $20$  percent of success.

The test mean and the percent successes indicate that this was a very difficult test. However, the test seemed to offer considerable promise since the reliability of the test was high and the items were on the average quite discriminating. Table IX presents a summary of the pertinent data for Test G.

TABLE IX

## PERTINENT DATA FOR TEST G

Number of items .....	100
Range of scores .....	6 - 64
Mean .....	$38.01 \pm 0.92$
Standard deviation .....	$11.95 \pm 0.65$
Reliability coefficient .....	$.90 \pm .01$
Range of discrimination indices .....	$-4 - 80$
Mean discrimination index .....	$31.82$
Range of difficulty indices .....	$0 - 76$
Mean difficulty index .....	$32.54$

Analysis of Test H - Interpretation of Data. Test H and Test J were presented to the students as a single test of  $168$  items (see page  $135$ , Chapter IV). However, for the purposes of analysis this single test was considered as two tests; Test H, Interpretation of Data and Test J,

Generalizations and Assumptions. The 75 items of the 168 item test which were answered by the key: true, probably true, insufficient data, probably false, and false, constituted Test H. The range of scores for this test was from 16 to 48. The mean of the test was  $32.19 \pm 0.49$  and the standard deviation was  $6.38 \pm 0.35$ . The estimated reliability was  $.70 \pm .04$ .

Complete item analysis data on Test H are presented in Table XXXXIII of Appendix I. The range of item discriminations expressed as an estimated coefficient of correlation with the total test score was from  $-.27$  to  $.76$ . The discrimination indices ranged from  $-17$  to  $60$ , resulting in a mean of  $24.69$ . This corresponds to an estimated coefficient of correlation with the total test score of  $.39$ .

The range of item difficulties was from 0 to 89 percent of successes. The range of indices of difficulty was from 0 to 76 giving a mean difficulty index of  $35.69$  and 25 percent success on the items. This figure and the mean of the test gave evidence that the test as a whole was quite difficult. A summary of the pertinent data for Test H is given in Table X.

TABLE X

## PERTINENT DATA FOR TEST H

Number of items .....	75
Range of scores .....	16 - 48
Mean .....	$32.19 \pm 0.49$
Standard deviation .....	$6.38 \pm 0.35$
Reliability coefficient .....	$.70 \pm .04$
Range of discrimination indices .....	$-17 - 60$
Mean discrimination index .....	24.69
Range of difficulty indices .....	0 - 76
Mean difficulty index .....	35.69

Analysis of Test J - Generalizations and Assumptions.

Test J, consisting of 93 items of the 168 items which constituted the combination Tests H and J, was designed to measure an understanding of generalizations and assumptions. The scores on this test ranged from 16 to 59. The mean of Test J was  $37.37 \pm 0.71$  while the standard deviation was  $9.31 \pm 0.51$  and the estimated reliability of the test was  $.81 \pm .03$ .

Complete item analysis data for Test J are presented in Table XXXIV of Appendix I. The range of item validity values was from  $-.04$  to  $.81$ . The discrimination indices ranged from 0 to 68. The mean discrimination index was 25.76. This is equivalent to an estimated coefficient of correlation of  $.40$ . The item difficulties ranged from 0 to 66 in terms of percents answering the item correctly. The range of difficulty indices was 0 to 59 and the mean was 34.62. This figure corresponds to a value of 23 percent when converted into percent passing the item.

The mean of the test and mean item difficulty both testified that this test, like Test H, was quite difficult. Table XI presents a summary of the pertinent data for Test J.

TABLE XI

PERTINENT DATA FOR TEST J

Number of items .....	93
Range of scores .....	16 - 59
Mean .....	$37.37 \pm 0.71$
Standard deviation .....	$9.31 \pm 0.51$
Reliability coefficient .....	$.81 \pm .03$
Range of discrimination indices .....	0 - 68
Mean discrimination index .....	25.76
Range of difficulty indices .....	0 - 59
Mean difficulty index .....	34.62

The data on the means, standard deviations, and reliabilities for all of the tests of the tryout battery are summarized in Table XII. The two least reliable tests were (1) Test B, which purported to measure the ability to delimit problems and (2) Test C, which was designed to measure an understanding of experimental design. Test D was the most reliable. This test, designed to measure ability to organize data, contained items which probably tested a very narrow range of ability and items which were all very similar. Test A, purporting to measure knowledge of steps of scientific thinking, Test F, designed to measure ability to interpret data and an understanding of controls, and Test G, designed to measure ability to draw conclusions, were all fairly reliable.

TABLE XII

COMPARISON OF MEANS, STANDARD DEVIATIONS,  
AND RELIABILITIES OF THE TRYOUT TESTS

Test	No. of Items	Mean	Standard Deviation	Reliability
A	74	50.60 $\pm$ .62	8.13 $\pm$ .44	.87 $\pm$ .02
B	50	22.46 $\pm$ .37	4.77 $\pm$ .26	.61 $\pm$ .05
C	62	26.30 $\pm$ .41	5.31 $\pm$ .29	.59 $\pm$ .05
D	20	10.94 $\pm$ .32	4.12 $\pm$ .23	.93 $\pm$ .01
E	74	34.37 $\pm$ .49	6.38 $\pm$ .35	.71 $\pm$ .04
F	72	47.85 $\pm$ .66	8.48 $\pm$ .46	.89 $\pm$ .02
G	100	38.01 $\pm$ .92	11.95 $\pm$ .65	.90 $\pm$ .01
H	75	32.19 $\pm$ .49	6.38 $\pm$ .35	.70 $\pm$ .04
J	93	37.37 $\pm$ .71	9.31 $\pm$ .51	.81 $\pm$ .03

A summary of item analysis data for all of the tests of the tryout battery is presented in Table XIII. Inspection of this table reveals that the mean item discrimination indices were all above the criterion value of 20 suggested by Davis.<sup>8</sup> Test D, the test to measure ability to organize data, had the highest mean discrimination index of any of the tests. Tests A, F, and G, judged on the basis of mean discrimination indices, were the next most successful tests. Test C, judged on the same basis, was the poorest. It is of interest to note that the rank order of the mean discrimination indices is very similar to the rank order of the reliabilities of the tests.

TABLE XIII

COMPARISON OF MEAN ITEM VALIDITIES AND  
MEAN ITEM DIFFICULTIES OF THE TRYOUT TESTS

Test	Mean Discrimination Coefficient	Mean Discrimination Index	Mean Percent Success	Mean Difficulty Index
A	.45	29.45	60	55.51
B	.44	27.08	30	39.40
C	.34	21.52	23	34.37
D	.70	52.60	42	45.90
E	.38	24.60	32	40.57
F	.47	30.66	59	55.13
G	.48	31.82	20	32.54
H	.39	24.69	25	35.69
J	.40	25.76	23	34.62

<sup>8</sup> Davis, op. cit., p. 15.

The mean difficulty indices ranged from 32.54 to 55.51, indicating that the tests were all relatively difficult. A criticism of the tests as a whole might be that they were a little too difficult for the group for which they were intended.

Analysis of tryout tests considered as a single test.

In all there were 620 items used in the determination of the scores on the total tryout battery. The range of scores was from 183 to 399. The mean for the entire battery of tests was  $291.12 \pm 3.48$ , while the standard deviation was  $44.22 \pm 2.26$ . The minimum reliability of the test, as estimated by the Kuder-Richardson<sup>9</sup> formula, was  $.92 \pm .01$  for this group of students. Table XIV presents a summary of the pertinent data for the tryout test battery.

TABLE XIV

PERTINENT DATA FOR THE TRYOUT TEST BATTERY

Number of items .....	640
Range of scores .....	183 - 399
Mean .....	$291.12 \pm 3.48$
Standard deviation .....	$44.22 \pm 2.26$
Reliability coefficient .....	$.92 \pm .01$

Intercorrelation of tryout test scores. In order to determine whether there was sufficient overlapping in the tests to justify the elimination of any of the types of items presented in the tryout tests in the preparation of the final form of the test, intercorrelations were calculated for all of the tryout tests. These intercorrelations are presented

---

<sup>9</sup> Adkins, op. cit., p. 154.

in Table XV. The standard errors of these correlations were small; they ranged from .05 to .08.

TABLE XV  
INTERCORRELATIONS OF TRYOUT TEST SCORES

Tests	Tests								
	A	B	C	D	E	F	G	H	J
A*		.18	.34	.27	.28	.37	.34	.44	.44
B			.21	.22	.30	.32	.18	.16	.11
C				.22	.26	.39	.32	.35	.33
D					.26	.26	.28	.29	.14
E						.47	.50	.45	.41
F							.47	.47	.45
G								.50	.31
H									.59
J									

\* A, Steps in Scientific Thinking. B, Delimitation of Problems. C, Experimental Procedures. D, Organization of Data. E, Evaluation of Hypotheses. F, Experimentation and the Interpretation of Data. G, Drawing of Conclusions. H, Interpretation of Data. J, Generalizations and Assumptions.

These data show that Test D, the test devised to measure ability to organize data, had a low correlation with all of the other tests of the battery. Tests H and J, the tests devised to measure interpretation of data, and the ability to recognize generalizations and assumptions respectively, which were presented to the students as a single test, had the highest intercorrelation of any of the tests. Was this due to



the fact that the same subject matter was used for both tests? Or was it due to the fact that an understanding of generalizations and assumptions was necessary for correct interpretation of data? The data presented are not such that they suggest possible answers to these questions.

The correlation between two tests is considered to be lowered if the test scores are unreliable.<sup>10</sup> In order to estimate the correlation between the true scores of two tests a correction known as the correction for attenuation<sup>11</sup> is frequently made which takes the unreliability of both tests into account. This correction gives the maximum correlation which could be obtained between the two test scores if both measures were perfectly reliable; that is, if the reliability coefficient of each test was 1.00. It must be kept in mind, however, that this is a theoretical value. The inter-correlations corrected for attenuation are given in Table XVI.

A comparison of Tables XV and XVI reveals the fact that all of the correlations have been increased by the correction for attenuation. The comparison also shows that the corrections of tests which were quite reliable, as Test D, were increased much less than tests which were rather unreliable, like Tests B and C. In addition, it can be seen that the lower correlations were increased less than the higher

---

<sup>10</sup> Henry E. Garrett, Statistics in Psychology and Education. New York: Longmans, Green and Company. 1947. p. 396.

<sup>11</sup> Loc. cit.

correlations.

TABLE XVI  
INTERCORRELATIONS OF TRYOUT TEST SCORES  
CORRECTED FOR ATTENUATION

	Tests								
Tests	A	B	C	D	E	F	G	H	J
A*		.25	.48	.30	.35	.42	.39	.56	.52
B			.35	.29	.45	.44	.24	.24	.17
C				.30	.40	.53	.44	.55	.48
D					.32	.29	.30	.36	.16
E						.59	.63	.63	.54
F							.53	.59	.53
G								.63	.36
H									.73
J									

\* A, Steps in Scientific Thinking. B, Delimitation of Problems. C, Experimental Procedures. D, Organization of Data. E, Evaluation of Hypotheses. F, Experimentation and the Interpretation of Data. G, Drawing of Conclusions. H, Interpretation of Data. J, Generalizations and Assumptions.

Since the purpose of these correlations was to determine whether there was sufficient overlapping of factors in the tests to warrant the omission of certain of these types of items in the preparation of the final form of the test, the degree of overlapping was determined by the coefficient of determination.<sup>12</sup> This figure is obtained by squaring the

<sup>12</sup> Ibid., p. 338.

coefficient of correlation. In order to obtain a figure representing the maximum overlap, the coefficients of correlation corrected for attenuation were used. The coefficient of determination denotes the percent of variance in one test associated with the other test. This figure is usually expressed as a percent. For example, the coefficient of determination between Test A and Test B is .06, which means that 6 percent of the variance of Test A is associated with Test B. The coefficients of determination for the tryout tests are given in Table XVII.

TABLE XVII

## COEFFICIENTS OF DETERMINATION OF THE TRYOUT TESTS

Tests	Tests								
	A	B	C	D	E	F	G	H	J
A*		.06	.23	.09	.12	.18	.15	.31	.27
B			.12	.08	.20	.19	.06	.06	.03
C				.09	.16	.28	.19	.30	.23
D					.10	.08	.09	.13	.03
E						.35	.40	.40	.29
F							.28	.35	.28
G								.40	.13
H									.53
J									

\* A, Steps in Scientific Thinking. B, Delimitation of Problems. C, Experimental Procedures. D, Organization of Data. E, Evaluation of Hypotheses. F, Experimentation and the Interpretation of Data. G, Drawing of Conclusions. H, Interpretation of Data. J, Generalizations and Assumptions.

The coefficients indicate that the degree of overlapping in these tests is low. Since the maximum overlapping is only 53 percent all of the types of items represented in the tryout battery were used in the construction of Test I, The Ability to Think Scientifically.

To determine whether the correlation between any of the tests of the tryout test battery of tests was sufficiently high to be used instead of the composite of the scores on the tryout test battery, the scores on each of the tests were correlated with the total scores. These correlations are given in Table XVIII.

TABLE XVIII

CORRELATION OF TOTAL SCORES ON TRYOUT TEST  
BATTERY WITH EACH OF THE TRYOUT TESTS

	Tests								
	A	B	C	D	E	F	G	H	J
Tryout total	.62	.44	.55	.41	.73	.74	.71	.71	.69

The standard errors of these coefficients ranged from .04 to .07. It is of interest to note that Test D, The Ability to Organize Data, had the lowest correlation with the total scores on the tryout test battery. This was to be expected on the basis of the nature of the test. Inspection of the test reveals that it was testing a much more restricted range of objectives than any of the other tests, therefore, it would not be expected that it would have as high a

correlation with the composite score as a test measuring a wider range of behaviors. That Test F, Experimentation and the Interpretation of Data, would have the highest correlation with the scores on the total test battery was to be expected, since that test measured both understanding of experimentation and the ability to interpret data, that is, it measured a wider range of the behaviors measured by the battery of tests than did any other tryout test.

Multiple correlations of the scores on the total tryout test battery with each combination of two of the individual tests of the battery were calculated to determine which two tests would be the most satisfactory to use in appraising the ability to think scientifically. The following formula<sup>13</sup> was used for the calculation of these multiple correlations:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 - 2r_{12}r_{13}r_{23} + r_{13}^2}{1 - r_{23}^2}}$$

---

<sup>13</sup> William D. Baten, Elementary Mathematical Statistics. New York: John Wiley and Sons. 1938. p. 187.

TABLE XIX

MULTIPLE CORRELATION OF TRYOUT TOTAL  
WITH TWO OF THE TRYOUT TESTS

Tests	Tests								
	A	B	C	D	E	F	G	H	J
A*		.69**	.72	.67	.84	.83	.82	.79	.77
B			.64	.54	.77	.77	.78	.78	.78
C				.63	.82	.81	.81	.78	.78
D					.76	.77	.74	.74	.76
E						.86	.83	.84	.85
F							.85	.85	.84
G								.82	.87
H									.77
J									

\* A, Steps in Scientific Thinking. B, Delimitation of Problems. C, Experimental Procedures. D, Organization of Data. E, Evaluation of Hypotheses. F, Experimentation and the Interpretation of Data. G, Drawing of Conclusions. H, Interpretation of Data. J, Generalizations and Assumptions.

\*\* This is to be read: Multiple correlation of tryout total with Tests A and B.

Table XIX is significant in that it shows that any two tests of the battery gave fairly substantial correlation with the criterion. Multiple correlations involving Test D were lower than any of the other correlations. The highest multiple correlation was obtained with Tests G and J. This is interesting since neither of these tests had the highest correlation with the criterion. This can probably be explained by the fact that they had a relatively low

correlation with each other as can be seen in Table XV.

In problems involving more than four variables the mechanics of calculating multiple correlations is almost prohibitive unless some systematic method of solution is used.<sup>14</sup> The Wherry-Doolittle method,<sup>15</sup> in addition to being a systematic method of calculating multiple correlations, corrects the correlation for chance errors. Table XX presents the results of this method of obtaining multiple correlations of the tryout tests with the criterion, which was the total score on the tryout test, and shows the correlations obtained by the addition of each successive test. In using this method the first test used, Test F, is the one with the highest simple correlation with the criterion.

TABLE XX  
MULTIPLE CORRELATION OF TRYOUT TESTS  
WITH THE CRITERION - OBTAINED BY  
THE WHERRY-DOOLITTLE METHOD

Tests	Multiple correlations
F	.740*
F, E	.856
F, E, J	.907
F, E, J, G	.948
F, E, J, G, B	.963
F, E, J, G, B, C	.972
F, E, J, G, B, C, H	.977

\* A simple correlation

<sup>14</sup> Garrett, op. cit., p. 435.

<sup>15</sup> Ibid., pp. 435 - 448.

Tests A and D were not added because the increase in the multiple correlation by the addition of Test H had been so slight that further additions seemed unnecessary. As shown by the data presented in Table XX each successive test added less to the correlation. It would appear that if a few of the individual tests of the tryout battery were to be used as a measure of the ability to think scientifically, Test E, The Evaluation of Hypotheses, Test F, Experimentation and the Interpretation of Data, Test J, Generalizations and Assumptions, and Test G, Drawing of Conclusions, would yield scores sufficiently like the ones obtained from the entire battery to justify the use of only these four tests.

Correlations of scores on tryout tests with scores on intelligence tests and reading tests. In order to determine whether the tryout tests were measuring intelligence or reading ability to a considerable extent, the scores made by students on each of the tryout tests were correlated with the quantitative score and with the linguistic score on the American Council on Education Psychological Examination and with the scores on the American Council on Education Reading Comprehension Test. These correlations are presented in the correlations of tryout test scores with intelligence test and reading test scores in Table XXI.



TABLE XXI

CORRELATIONS OF TRYOUT TEST SCORES WITH  
INTELLIGENCE TEST AND READING TEST SCORES

Tests	Tests		
	Quantitative	Linguistic	Reading
A*	.17	.43	.25
B	.24	.26	.13
C	.28	.41	.39
D	.31	.11	.10
E	.35	.42	.41
F	.38	.34	.35
G	.37	.21	.29
H	.37	.18	.25
J	.37	.29	.35

\* A, Steps in Scientific Thinking. B, Delimitation of Problems. C, Experimental Procedures. D, Organization of Data. E, Evaluation of Hypotheses. F, Experimentation and the Interpretation of Data. G, Drawing of Conclusions. H, Interpretation of Data. J, Generalizations and Assumptions.

The abilities measured by the tryout tests, although all positively correlated with the quantitative and linguistic factors of intelligence and with reading ability, do not appear to be identical with any of these mental functions. These data also give evidence that the inclusion of all of the types of items presented in the tryout battery could justifiably be included in the final form of the test since none of the tryout tests seemed to be measuring either of the factors of intelligence or reading ability.

The preparation of Test I. - The Ability to Think Scientifically. Test I, presented in Appendix II, was constructed from items of the tryout tests. Because of the nature of the items it was necessary to choose blocks of items from the tryout tests rather than individual items. Therefore, it was necessary to select the best blocks of items from each of the tryout tests. Items within blocks were eliminated if the estimated coefficient of correlation of the item with the tryout test score was low. Certain items had to be retained even if the item correlation was low because the information given in them was essential to the development of an understanding of the entire block of items. An attempt was made to eliminate all items with a discrimination index of less than 20 which corresponds to a coefficient of correlation of .33. This is in accord with the recommendation of Davis.<sup>16</sup>

Test authorities do not agree on the best form of distribution of item difficulties.<sup>17</sup> Some recommend all items as near 50 percent difficulty as possible; others recommend equal distribution of items from 0 to 100 percent difficulty.

---

<sup>16</sup> Davis, op. cit., p. 15.

<sup>17</sup> Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, The Construction and Use of Achievement Examinations. Cambridge: Houghton Mifflin Company. 1936. p. 32.

Flanagan<sup>18</sup> has shown that on a theoretical basis the best test would be one composed of items which were all answered correctly by 50 percent of the group if the correlations between individual items were zero; but that in a theoretical case where the correlations between the individual items of the test were one, the items should range from zero to 100 percent of successes. Both of these situations are, of course, hypothetical. In reality, the situation is usually intermediate between these two extremes. In the present case items have been chosen from a range of from 10 - 95 percent difficulty. This is in accordance with the suggestion of Hawkes, Lindquist and Mann.<sup>19</sup>

Ten to 20 items were selected from each of the tryout tests with the exception that only four of the best items were selected from Test D. Test I, The Ability to Think Scientifically (see Appendix II), was made up of 150 items selected from a total of 637 items which comprised the tryout test battery. An attempt was made to include items to appraise most of the behaviors identified in Chapter IV, therefore only four items were used from Test D, despite the high discrimination index of the items.

---

<sup>18</sup> Flanagan, op. cit., p. 675-676.

<sup>19</sup> Hawkes, Lindquist, and Mann, op. cit., p. 32.

## ANALYSES OF TEST I AND TEST IA

Test I, The Ability to Think Scientifically, was administered to 500 students who had completed a year of Biological Science at the end of the spring term of 1950 and to 240 students who had not yet had Biological Science at the beginning of the fall term of 1950. Test IA, the final form of the test, The Ability to Think Scientifically, was administered to 330 other students who had not yet taken Biological Science at the beginning of the fall term of 1950. Test IA was also administered to 136 of this same group at the end of the first term of the course.

Analysis of Test I - The Ability to Think Scientifically. Test I was comprised of a total of 150 items. The range of scores for the students who had completed the year course was from 30 to 117. The mean of the scores was  $78.92 \pm .73$ ; the standard deviation was  $15.41 \pm .52$ . The reliability of the test for this group was  $.89 \pm .01$  as determined by the split-half method, corrected by the Spearman-Brown formula. By using the Kuder-Richardson formula a reliability of  $.85 \pm .01$  was obtained.

The range of scores for the students who had had no college biology was from 27 to 107. The mean for this group was  $60.64 \pm 1.18$ , the standard deviation was  $17.32 \pm .83$ . The reliability for this group was  $.91 \pm .01$  calculated by the split-half method, corrected by the Spearman-Brown

formula. The Kuder-Richardson formula gave a value of  $.89 \pm .01$ .

The complete item analysis data for Test I are presented in Table XXXXV of Appendix II. The papers of the 500 students who completed this test in the spring term of 1950 were used for the analysis. Item discriminations were determined by the methods of Flanagan<sup>20</sup> and of Davis<sup>21</sup> as described previously. The criterion used was the total score on Test I, The Ability to Think Scientifically.

The item discrimination values ranged from  $-.23$  to  $.73$ . The value of  $-.23$  is somewhat difficult to explain since discriminating items from the tryout test had been used in the construction of Test I. The discrimination indices ranged from  $-14$  to  $56$ , giving a mean discrimination index of  $25.36$ . This corresponds to a value of  $.39$  expressed as an estimated coefficient of correlation with the total test score. Item difficulties were estimated by the methods suggested by Davis.<sup>22</sup> The difficulties expressed in percent of successes ranged from  $0$  to  $86$  percent. The indices of difficulty ranged from  $0$  to  $73$ , with an average difficulty index of  $43.17$ . This corresponds to an average of  $32$

---

<sup>20</sup> Flanagan, op. cit., pp. 674-680.

<sup>21</sup> Davis, op. cit., pp. 8-15.

<sup>22</sup> Ibid., pp. 2-4.

percent successes.

The mean of the test and the mean item difficulty gave evidence that the test was probably a little too difficult for the group for which it was devised. The reliability of the test compares favorably with the requirements of most standardized tests. The pertinent data on Test I, The Ability to Think Scientifically, are presented in Table XXII.

TABLE XXII  
PERTINENT DATA FOR TEST I

	Group	
	3 terms Biological Science	No Biological Science
Number of students .....	500	240
Number of items .....	150	150
Range of scores .....	30 - 117	27 - 107
Mean .....	78.92 $\pm$ .73	60.64 $\pm$ 1.18
Standard deviation .....	15.41 $\pm$ .52	17.32 $\pm$ .83
Reliability coefficient .....	.89 $\pm$ .01	.91 $\pm$ .01
Range of discrimination indices	-14 - 56	
Mean discrimination index ....	25.36	
Range of difficulty indices ..	0 - 73	
Mean difficulty index .....	43.17	

A comparison of the discrimination indices and the difficulty indices as determined for the same items in the tryout tests and Test I is presented in Table XXIII. The data in this table constitute evidence that the test items chosen from the tryout tests to make up Test I were, in general, highly discriminating.

TABLE XXIII

COMPARISON OF DISCRIMINATION INDICES AND  
OF DIFFICULTY INDICES OF IDENTICAL ITEMS  
AS OBTAINED FROM ITEM ANALYSIS OF TRYOUT TESTS  
AND AS OBTAINED FROM ITEM ANALYSIS OF TEST I

<u>Item Number</u>		<u>Discrimination Index</u>		<u>Difficulty Index</u>	
Tryout Test	Test I	Tryout Test	Test I	Tryout Test	Test I
A-13	1	15	10	72	73
A-14	2	40	15	53	55
A-15	3	43	20	53	55
A-16	4	25	33	70	68
A-17	5	29	34	68	68
A-18	6	30	17	47	40
A-19	7	40	23	73	60
A-20	8	27	10	69	59
A-21	*9	19	- 6	54	45
A-22	10	29	28	68	53
A-23	11	49	42	32	27
A-24	*12	16	- 2	46	40
A-25	13	12	26	56	60
A-26	14	40	28	38	44
A-27	15	33	17	62	55
A-28	16	28	23	55	51
A-29	17	34	14	56	52
B- 1	*18	18	20	38	43
B- 5	*19	48	37	45	29
B- 7	*20	43	-14	28	9
B- 8	*21	25	8	34	37
B- 9	*22	48	32	42	39
B-10	*23	22	10	55	52
B-11	*24	29	10	52	52
B-12	*25	16	23	49	46
B-13	*26	45	17	43	47
B-14	*27	34	22	53	54
B-19	*28	24	16	33	37
B-21	*29	72	33	44	35
B-22	*30	13	14	49	45
B-25	*31	46	27	30	36
B-28	*32	39	29	46	36
B-30	*33	29	12	38	45
C- 3	34	63	37	40	25
C- 6	*35	58	1	38	1
C- 7	36	38	50	43	33
C- 8	37	42	26	66	59
C-21	38	31	35	51	50
C-23	39	23	32	61	55

\* Items eliminated from Test I in construction of Test IA

TABLE XXIII (continued)

<u>Item Number</u>		<u>Discrimination Index</u>		<u>Difficulty Index</u>	
Tryout Test	Test I	Tryout Test	Test I	Tryout Test	Test I
C-26	40	31	27	44	40
C-28	*41	20	2	45	42
C-29	42	36	11	54	56
C-51	43	54	22	40	42
C-55	*44	16	46	34	30
C-56	*45	52	3	34	38
C-58	*46	41	0	34	0
C-62	47	34	24	53	46
D-13	48	90	24	48	46
D- 8	49	71	20	51	56
D-16	50	58	13	38	45
D- 5	51	28	9	55	56
E- 1	52	31	14	42	32
E- 4	53	51	50	33	33
E- 5	54	26	22	51	51
E- 6	55	42	23	34	34
E- 7	56	29	20	64	63
E- 8	57	9	16	70	67
E- 9	58	37	16	52	44
E-10	59	11	8	44	45
E-11	60	15	21	41	39
E-12	61	38	48	32	38
E-47	62	40	19	38	45
E-49	63	38	27	60	60
E-50	64	51	33	33	36
E-51	65	23	23	61	47
E-52	66	45	31	36	20
E-53	67	18	15	26	15
E-54	68	54	3	35	22
E-55	69	34	27	38	35
E-57	70	20	8	35	39
E-60	71	23	11	44	46
F-58	72	25	23	66	55
F-59	73	60	56	56	40
F-60	74	49	24	58	47
F-61	75	50	34	45	22
F-62	76	56	35	54	49
F-63	77	55	37	53	46
F-64	78	59	35	56	47
F-71	79	33	34	78	68
F-72	80	25	29	63	60
F-40	81	49	46	53	55
F-41	82	36	43	54	49



TABLE XXIII (continued)

<u>Item Number</u>		<u>Discrimination Index</u>		<u>Difficulty Index</u>	
<u>Tryout</u> <u>Test</u>	<u>Test I</u>	<u>Tryout</u> <u>Test</u>	<u>Test I</u>	<u>Tryout</u> <u>Test</u>	<u>Test I</u>
F-42	83	21	29	44	44
F-43	84	54	44	64	60
F-44	85	51	36	54	52
F-45	86	30	26	74	65
F-46	87	45	33	46	33
F-52	88	41	22	54	44
F-53	89	56	32	37	21
F-54	90	50	46	61	57
F-55	91	36	37	58	52
G- 1	92	28	27	41	40
G- 4	93	59	30	53	48
G- 5	94	33	22	50	49
G-15	95	47	44	30	33
G-17	96	33	32	66	59
G-18	97	47	20	30	13
G-20	98	49	31	68	61
G-35	99	23	34	48	55
G-39	100	82	41	47	44
G-40	101	43	36	40	33
G-41	102	52	48	34	31
G-47	103	0	0	0	0
G-48	104	68	47	42	30
G-50	105	30	26	64	57
G-51	106	56	40	37	33
G-53	107	49	25	32	16
G-54	108	71	22	51	48
G-89	109	28	29	49	43
G-90	*110	35	31	23	20
H-42	111	21	15	65	62
H-44	112	47	40	42	41
H-46	113	72	39	44	42
H-47	114	25	37	44	49
H-48	115	36	21	76	69
H-49	116	37	26	55	57
H-53	117	26	30	62	64
H-55	118	24	6	70	49
H-59	119	35	29	42	44
H-61	120	16	16	47	37
J-63	121	23	21	42	51
J-64	122	26	20	45	58
J-67	123	43	19	52	44
J-68	124	34	27	53	52
J-70	125	19	12	54	56

TABLE XXIII (continued)

<u>Item Number</u>		<u>Discrimination Index</u>		<u>Difficulty Index</u>	
<u>Tryout</u> <u>Test</u>	<u>Test I</u>	<u>Tryout</u> <u>Test</u>	<u>Test I</u>	<u>Tryout</u> <u>Test</u>	<u>Test I</u>
J-71	*126	33	14	35	9
J-74	127	63	27	40	43
J-75	128	32	10	54	45
J-77	129	50	39	38	42
J-78	130	53	30	35	33
J-80	131	40	27	41	45
H-84	132	44	25	46	40
H-85	133	31	32	66	59
H-86	134	36	16	37	23
H-88	135	47	19	42	34
H-91	136	28	47	56	42
H-93	137	35	22	69	62
H-95	138	27	35	42	44
H-99	139	46	16	30	10
H-100	140	24	5	48	48
J-101	141	59	49	47	36
J-104	142	56	24	36	31
J-105	143	28	50	50	47
J-106	144	45	37	51	49
J-110	145	31	21	46	52
J-111	146	32	38	40	25
J-116	147	52	43	34	32
J-119	148	29	16	36	10
J-122	149	55	35	36	27
J-123	150	30	22	53	47

\* Items eliminated from Test I in construction of Test IA

A few items of low discrimination were included because these items gave information necessary to the answering of subsequent items. This was especially true of item 103 which had no discriminative value. Omitting this single item the range of discrimination indices was from 9 to 90, which corresponds to a range of from .18 to .90 when expressed as an estimated coefficient of correlation with the criteria (the individual tryout test scores). The average discrimination index of the items, based on tryout data, was 37.90. This corresponds to an estimated coefficient of correlation of .56. This value represents relatively high item validity.

It is of interest to note that the discrimination indices obtained by using Test I as the criterion are, with a few exceptions, lower than the indices obtained by using the individual tryout tests as the criterion. Since each of the tryout tests was constructed to measure a rather narrow range of abilities, individual items would be expected to be more highly correlated with the score of the single tryout test, of which they were a part, than with the test on many of the abilities involved in scientific thinking. In other words, Test I had less internal consistency than the individual tryout test.

Table XXIV presents a comparison of (1) the item analysis data of the items of the tryout tests used in the construction of Test I, (2) the item analysis data on Test I

using the total score on this test as the criterion and (3) the item analysis data on the items of Test I used in the construction of Test IA. (Item 103 has been omitted from these comparisons since it had been included because it was necessary to the development of the idea presented in the block of items of which it was a part).

TABLE XXIV

SUMMARY OF ITEM ANALYSIS DATA FOR TRYOUT TEST ITEMS USED IN CONSTRUCTION OF TEST I, ITEMS OF TEST I, AND ITEMS OF TEST I USED IN CONSTRUCTION OF TEST IA

	Tryout	Test I	Test IA
Number of items .....	150	150	125
Range of discrimination indices	9 - 90	-14 - 56	3 - 36
Mean discrimination index ....	37.90	25.36	27.22
Range of difficulty indices ..	26 - 78	0 - 73	10 - 73
Mean difficulty index .....	47.82	43.17	44.64

Analysis of Test IA - The Ability to Think Scientifically. Since items were presented in blocks centering around a particular problem or experiment, they could not be arranged in order of difficulty. It was not intended that the test be designed as a speed test, and the nature of the sequence of items was not such that it could be arranged as a power test. Since the test was devised to measure growth and was to be used as a means of evaluating instruction, it seemed advisable to make the test of such a length that all, or at least 99 percent, of the students could finish it in the allotted time. As reported in

Chapter III, 10 percent of the students failed to complete Test I in the hour and fifty minutes available, therefore 25 of the poorer items of Test I were eliminated on the basis of item analysis to make Test IA. These included the 16 items of the section on the test designed to measure the student's ability to delimit a problem. This portion of the test had not proved satisfactory. The key, that is the five choices available for answering the items, was probably not satisfactory since there seemed to be too much overlapping. It was difficult for the student to determine whether a question was or was not related to the problem and whether, if related, it might or might not be useful in the solution of the problem. It seemed to the writer that this type of item had promise for future tests with revision of the possible answers, but because these items did not seem to contribute to the test as a whole, it seemed advisable to eliminate them. Test IA is presented in Appendix III.

The items marked with an asterisk in Table XXIII are the ones which were dropped from Test I in the construction of Test IA. Test IA was administered at the beginning of the fall term to 330 students who had had no Biological Science. The range of scores made by this group was from 23 to 101. The mean of Test IA was  $53.16 \pm 1.11$ , while the standard deviation was  $16.28 \pm .78$ . The reliability as determined by the split-half method with the Spearman-Brown correction was  $.91 \pm .01$ . The Kuder-Richardson formula gave a reliability of  $.89 \pm .01$ .

This test was administered to 136 of the same students at the end of the fall term of 1950 after they had taken one term of Biological Science. The range of scores for this group was from 31 to 103. The mean was  $69.94 \pm 1.41$ , and the standard deviation  $16.43 \pm .99$ . The reliability as determined by the split-half method corrected by the Spearman-Brown formula was  $.90 \pm .02$ . The reliability as calculated by the Kuder-Richardson formula was  $.89 \pm .02$ . Table XXV presents the pertinent data for Test IA.

TABLE XXV  
PERTINENT DATA FOR TEST IA

	Group	
	No Biological Science	136 of same group after one term of Biological Science
Number of students .....	330	136
Number of items .....	125	125
Range of scores .....	23 - 101	31 - 103
Mean .....	$53.16 \pm 1.11$	$69.94 \pm 1.41$
Standard deviation .....	$16.28 \pm 0.78$	$16.43 \pm 0.99$
Reliability coefficient	$.91 \pm .01$	$.90 \pm .02$

Item analysis data were not collected for Test IA, since this test was constructed by omitting items from Test I. Table XXIV presents the range of discrimination indices, mean discrimination index, the range of difficulty indices, and the mean difficulty index for the items used in this test. As can be seen from an inspection of this table, both the mean discrimination index and the mean difficulty index

were higher than corresponding values for Test I. This probably accounts for the fact that the reliability of Test IA was at least as high as the reliability of Test I, even though it was 25 items shorter.

## CHAPTER VI

### THE VALIDATION OF THE TEST

This chapter is concerned with the validation of the test, The Ability to Think Scientifically. The methods used in the curricular validation and the methods used in the statistical validation are presented.

The most important characteristic of a test is its validity, that is, the extent to which a test measures what it is supposed to measure.<sup>1, 2</sup> Validity is not a general term which can be applied to a test, but is a very specific concept and must be considered with reference to the purpose for which the test is used. A test is valid only in so far as it accomplishes its specific purpose for a particular group.

Remmers and Gage<sup>3</sup> have discussed the kinds of criteria which have been used in the validation of tests. They divide the criteria into two interrelated classes; (1) criteria with

---

<sup>1</sup> Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, The Construction and use of Achievement Tests. Cambridge: Houghton Mifflin Company. 1936. p. 21

<sup>2</sup> Dorothy C. Adkins, Construction and Analysis of Achievement Tests. Washington: U. S. Government Printing Office. 1947. p. 160.

<sup>3</sup> Hermann H. Remmers and N. L. Gage, Educational Measurement and Evaluation. New York: Harper and Brothers. 1943. pp. 195-201.



which to compare test content, and (2) criteria with which to compare test scores.

They state that the criteria with which the content of a test may be compared are; (1) analysis of courses of study, (2) statements of instructional objectives, (3) analysis of text books, (4) analysis of teacher's final examination questions, (5) pooled judgments of competent persons, (6) concepts of social utility, and (7) introspective logical or psychological analysis of mental processes. These types of criteria have been referred to as curricular criteria.

The criteria which Remmers and Gage mention to which scores on the test may be compared are; (1) school marks, (2) increases in percentage of success in successive ages or grades, (3) differences in scores obtained by any two or more groups known to be widely separated in ability, (4) ratings of pupils by competent raters, and (5) correlations with other tests. The validity obtained by these methods has been referred to as statistical validity.

#### THE CURRICULAR VALIDATION OF THE TEST

The course of study for Biological Science at Michigan State College was analysed and objectives of the course were considered in the construction of the test. The curricular validity of the test was insured by the incorporation into the test of the desired educational outcomes related to

scientific thinking which are emphasized in the course. In addition, a detailed description of the behaviors involved in scientific thinking was undertaken. This detailed analysis, presented in Chapter IV, was based upon the analysis of behaviors involved in scientific thinking as (1) described by persons constructing tests designed to measure the ability to think scientifically, (2) inferred from the elements of scientific thinking, (3) described in committee reports on behaviors involved in scientific thinking, and (4) described in reports of research on behaviors of persons doing scientific research. In all 98 behaviors attending scientific thinking were outlined. Test items were constructed from the outline of behaviors presented in Chapter IV, and an attempt was made to include as many of these behaviors as possible in the tryout tests. An inspection of Table I, (see pages 138 - 142), indicates that most of the 98 abilities which were identified as critical aspects of scientific thinking were appraised by the tryout tests.

A number of tests designed to measure the abilities involved in problem-solving were examined. The analysis of these tests revealed the kinds of techniques which had been used to measure the ability to think scientifically, and thus provided a basis for the curricular validity of the test. The use of some of the techniques used previously and an attempt to include items in the tryout tests which measured most of

the behaviors measured by previous tests should contribute to the curricular validity of the test.

Another method consisted of submitting the tryout tests to five competent judges for criticism. The judges agreed that the items were valid measure of the abilities which each of the tryout tests purported to measure. In addition, there was substantial agreement as to the correct answer for each item. Where there were disagreements among the judges the items were discussed with each of them and these items were either revised on the basis of the discussion or were eliminated.

Free responses of students were used as items of the test whenever this method of obtaining items was feasible. Situations were presented to students and students were requested to indicate what problems were suggested by these situations. The problems suggested were utilized in the construction of the test devised to measure the ability to delimit problems. Hypotheses were presented; students were requested to describe experiments to test these hypotheses. These experiments were used in the construction of the test devised to measure the ability to plan experiments. Data were given to students; the students were instructed to draw conclusions from the data. These conclusions were used in the construction of the test to appraise the ability to draw conclusions. In all cases the groups from which free responses were obtained were different groups from those to

which the tryout tests were administered. The use of free responses of students should contribute to the validity of the test because items written by students should be comprehensible to other students and because the responses of the students represent the kinds of answers which students give on essay type examinations.

Careful selection of materials for the test items should also contribute to the validity of the test. The criteria used in the selection of materials were discussed in Chapter III. The ones of importance to the validation of the test were:

1. The material should be comprehensible to students who had had no training in biology.
2. Data used for interpretation should be entirely new to the student.
3. The material should be biological since the test was devised for a course in first year college biology.

The first of these criteria was met by the selection of materials which were on subject matter which it was assumed all students had encountered, such as colds, disease, breathing, plants, etc. The second criterion was met by choosing data from sources which the elementary student would not be expected to read, such as scientific journals and advanced text books. The third criterion was satisfied by using materials of a biological nature.

Perhaps the most important method of validating a test is by considering its social utility. The committee reports

reviewed in Chapter I testify to the fact that the ability to think scientifically is one of the important objectives of general education, and a primary objective of the teaching of science. The test, designed to measure an objective of importance in a democratic society should have social usefulness.

### THE STATISTICAL VALIDATION OF THE TEST

Validation by correlation with measures of intelligence, reading ability, and factual information. The first method used to establish the statistical validity of the test was the correlation of scores made by students on the tests with other kinds of tests. In a sense, this is a negative form of validation because a high correlation of this test with measures of such traits as intelligence, reading ability, and knowledge of facts would indicate that the test could not then measure in any considerable amount what it purported to measure, assuming that the test was designed to measure something different. It cannot be assumed, however, that the test is a valid measure of ability to think scientifically merely because of a lack of substantial relationship to any of these factors.

Preliminary evidence concerning the statistical validity of the test was obtained by correlating the total scores made by 162 students taking the tryout tests with; (1) the quantitative scores on the psychological examinations, (2)

the linguistic scores of the psychological examinations, (3) the total psychological examination scores, (4) total reading examination scores, (5) the sums of the scores made on the departmental term-end examinations for first and second terms, (6) the scores on the factual portion of the comprehensive examination, and (7) the scores on the portion of the comprehensive examination which involved the use of scientific thinking as well as biological information. It was assumed that a high correlation between the scores on the battery of tryout tests and the scientific thinking portion of the comprehensive examination would give some positive evidence of validity. It was also assumed that low correlations with the first six tests would be desirable.

As previously mentioned, psychological examination scores and reading examination scores were available for 264 of the 500 students who took Test I, The Ability to Think Scientifically, in May, 1950. The scores made by these 264 students on Test I were correlated with; (1) the quantitative scores on the psychological examination, (2) the linguistic scores on the psychological examination, (3) the total scores on the psychological examination, and (4) the total reading test scores. These correlations together with those obtained by correlating the same four factors with the total tryout test are given in Table XXVI. The standard errors of the correlations ranged from .04 to .07.

TABLE XXVI

CORRELATION OF TRYOUT TEST SCORES AND SCORES ON TEST I  
WITH PSYCHOLOGICAL EXAMINATION SCORES AND READING TEST SCORES

Tests	Tests			
	Quantitative	Linguistic	Total Psychological	Reading
Tryout	.45	.38	.51	.49
Test I	.48	.43	.51	.43

These data show that there was a moderate positive correlation between the ability to think scientifically, as measured by these tests, and both the quantitative and linguistic factors of intelligence, and reading ability.

Since the tryout tests and the American Council on Education Psychological Examination both depended to some extent on reading ability, it seemed desirable to hold reading ability as a constant factor in making a correlation between the ability to think scientifically and intelligence. This was accomplished by the use of a partial correlation. The formula<sup>4</sup> used in the calculation of the partial correlation was:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

The intercorrelations used to determine the partial correlation are given in Table XXVII.

<sup>4</sup> Quinn, McNemar. Psychological Statistics. New York: John Wiley and Sons. 1949. p. 141.

TABLE XXVII

INTERCORRELATIONS OF TRYOUT TEST,  
PSYCHOLOGICAL EXAMINATION AND READING TEST

Tests	Tests		
	Tryout	Psychological	Reading
Tryout		.51	.49
Psychological			.59
Reading			
Partial Correlation ( $r_{12.3}$ ) = .31			

The correlation between the ability to think scientifically, as measured by the battery of tryout tests, and intelligence, with reading ability held constant, was .31. This indicated that a part of the observed relationship between the ability to think scientifically and intelligence, as measured in these two tests, was due to the common dependence of both tests upon reading but that most of the relationship still remained.

A partial correlation was also calculated for the relationship between Test I, The Ability to Think Scientifically, and intelligence with reading ability held constant. The intercorrelations used in this computation are presented in Table XXVIII. The partial correlation was .33. These partial correlations give evidence that the ability to think scientifically was not identical to intelligence but that the abilities were related.



TABLE XXVIII  
INTERCORRELATIONS OF TEST I,  
PSYCHOLOGICAL EXAMINATION, AND READING TEST

Tests	Tests		
	Test I	Psychological	Reading
Test I		.51	.43
Psychological			.60
Reading			
Partial Correlation ( $r_{12.3}$ ) = .33			

In order to determine the degree to which reading ability and the ability to think scientifically were related, a partial correlation of the ability to think scientifically and reading ability, with intelligence held constant, was calculated. For the scores on the tryout test battery this partial correlation was .28; for the scores on Test I the partial correlation was .18. Both correlations are sufficiently low to show that the tests were not primarily reading tests.

In order to give a more complete picture of the relationship of the scores made on the tryout tests to various other abilities Table XXIX has been prepared. In this table are shown the correlations between (A) the total scores on the battery of tryout tests, (B) total scores on the American Council on Education Psychological Examination, (C) total scores on the American Council on Education Reading

Comprehensive Test, (D) total scores on departmental term-end examinations in Biological Science, (E) the scores on the factual portion of the comprehensive examination for Biological Science, and (F) the scores on the scientific method portion of the comprehensive examination for Biological Science.

TABLE XXIX

INTERCORRELATIONS OF TOTAL TRYOUT TEST SCORES  
AND SCORES ON OTHER TESTS

*	A	B	C	D	E	F
A	-	.51	.49	.65	.33	.70
B		-	.59	.52	.23	.58
C			-	.47	.36	.59
D				-	.52	.78
E					-	.41
F						-

\* A, Tryout Test Battery. B, A. C. E. Psychological Examination. C, A. C. E. Reading Test. D, Term-end examinations. E, Comprehensive, Fact. F, Comprehensive, Scientific method.

An inspection of Table XXIX reveals that the ability to think scientifically as measured by this battery of tests was positively related to all of the other factors measured by the other tests given.

A correlation of .70 (Table XXIX) obtained between the scores on the scientific method portion of the comprehensive

examination and the total scores on the tryout tests is evidence that the abilities involved in scientific thinking as defined by this investigator and the abilities involved in scientific thinking as defined by the trained examiner for the Department of Biological Science were in substantial agreement.

About 25 to 30 percent of the items on the departmental term-end examination are items designed to measure scientific thinking. On the basis of this fact one would expect a moderate degree of relationship between scores on the total tryout test and scores on the term-end examinations, however, the correlation of .65 may also indicate that there is a higher relationship between knowledge of facts and ability to think scientifically than the correlation between the tryout tests and scores on the factual portion of the comprehensive indicates. This correlation was .33. The relationship between knowledge of facts and ability to think scientifically should be further investigated.

Validation by comparison of scores of various groups.

Another method of statistical validation of the test was the comparison of scores made by students who had not yet taken Biological Science with scores made by students who had taken Biological Science.

The scores of students at the beginning of the course

in Biological Science were compared with the scores made by another group at the end of the three-term course in Biological Science. This comparison involved the assumption that the groups were both representative samples of the same population. In reality, this assumption is not strictly true since many persons proficient in Biological Science were permitted to take the comprehensive examination before completing three terms of the course, and hence were not represented in the group that had completed the three-term course in Biological Science. Also, more poor students drop out of school than good students, thus eliminating some of the lower scores. The lower standard deviation for this group gives evidence that these factors were operative. Equated groups might have been used to reduce the variability of the groups, but psychological examination scores and reading scores were not available at the time of administration of the test to the group beginning Biological Science.

The scores of students at the beginning of the course in Biological Science were also compared with their scores at the end of one term. This method relieves one of making an assumption concerning the nature of the group, but involves the assumption that memory would play no substantial part in any observed increase in scores. However, if the two methods gave substantially the same results valid inferences concerning the validity of the test could probably be

drawn. The validation of the test by these comparisons is based on the assumption that increase in scores results from instruction in the objective being tested and not on a maturation factor.

As previously mentioned, Test I was administered to 500 students who had completed three terms of Biological Science. Of this group 446 completed the test. The scores made by this group were compared with the scores made by 216 other students who completed the same test before taking Biological Science. A comparison of the scores of the two groups, as presented in Table XXX, gives evidence that there was improvement of scores and that this improvement was highly significant. The critical ratio of 13.15 showed that this difference between the two means was not due to chance.

TABLE XXX

COMPARISON OF MEANS AND STANDARD DEVIATIONS OF TEST I  
FOR A GROUP BEFORE TAKING BIOLOGICAL SCIENCE WITH ANOTHER  
GROUP AFTER TAKING THREE TERMS OF BIOLOGICAL SCIENCE

Group	Number	Mean	Standard Deviation
3 terms of Biological Science	446	78.92	15.41
No Biological Science	216	60.64	17.32
- - - - -			
$\frac{M_1 - M_2}{S.E. \cdot M_1 - M_2} = \text{Critical Ratio} = 13.15$			

A comparison was also made of the scores made by 136 of the group who took Test IA before taking Biological Science, that is, a pre-test group, and the scores on the same test made by this group after one term of Biological Science, a post-test group. The data for this phase of the study are presented in Table XXXI. The critical ratio of 8.62 gives further evidence that the difference between the two means was not due to chance.

TABLE XXXI  
COMPARISON OF MEANS AND STANDARD  
DEVIATIONS OF TEST IA ON THE PRE-TEST AND POST-TEST

Group	Number	Mean	Standard Deviation
Pre-test	136	55.60	15.84
Post-test	136	64.94	16.43
- - - - -			
$M_1 - M_2$ = Critical Ratio = 8.62			
S.E. $M_1 - M_2$			

The range of improvement on Test IA is of interest. Of the 136 students who retok the test, three did not change their scores, seven had scores from one to ten points lower on the post-test, and the remaining 126 students improved their scores from one to 41 points.

Since in both comparisons the differences between the means were highly significant and both in the same direction

we may make the inference that the test had some validity in that there was an increase in score attending instruction in the methods of science. One is obliged, however, to hold this inference as tentative until (1) further evidence concerning the relationship between increased knowledge of the subject matter of the course and performance on the test is further investigated, (2) until it is demonstrated that maturation did not produce the observed results, and (3) until it is shown that other methods of instruction do not produce the same results.

Validation by comparison of scores with ratings of students by competent judges. The final method used in the statistical validation of the test was the comparison of scores made on the test with the rating of competent judges. A rating scale for the ability to use the scientific method (Appendix IV) was prepared.

Several members of the Department of Biological Science at Michigan State College were interviewed in order to determine the types of behaviors which they had observed in students whom they considered to have superior ability to think scientifically and the types of behavior which they had observed in students whom they believed to be very inferior in this ability. The two areas in which they agreed that ratings of the students could be made on the basis of observation of their performance in laboratory classes were

(1) the ability to devise and evaluate experiments, and (2) the ability to interpret data, including the ability to form hypotheses and draw conclusions.

The instructions for rating the students were:

Will you please rate the person whose name appears above on the two following characteristics? The two extremes of these characteristics are described. Place a cross (X) on the line indicating your judgment of the individual with respect to the qualities in question.

A person having a high degree of ability to evaluate and devise experiments was described in the following manner:

Includes control factors, controls all but one variable, understands problem and devises experiment to test hypothesis. Can devise experiments which will yield results, recognizes problems inherent in the experiment, and has an understanding of what is happening in the experiment.

A person having a low degree of ability to evaluate and devise experiments was described:

Experiments lack control or control is faulty, experiment unrelated to hypothesis. Student does not understand the experimental set-up, or the problems inherent in the experiment.

Proficiency in ability to interpret data could be recognized by the following description of a person very superior in this ability:

Is able to make logical inferences from data, takes pertinent facts into consideration, applies previous knowledge to the new situation, is able to see relationships, especially cause and effect relationships. Knows what evidence for his inference is, and why it is evidence.



The person very inferior in this ability:

Is unable to make logical inferences from data, does not differentiate between relevant and irrelevant data or between critical and non-critical data, is unable to see relationships.

The ratings were on a five point scale; very superior, superior, average, inferior, and very inferior. One hundred and forty-three students taking the first term of Biological Science who were given Test IA at the beginning of the first term of the three-term sequence of the course were rated on their ability to think scientifically by their instructors. Test IA was administered again to 136 of these same students at the end of the first term. A part of these students were taught by the present investigator and the remaining students were taught by another instructor. Each of these students was rated by his instructor on the rating scale described above.

Students taking Biological Science at Michigan State College do not necessarily have the same instructor for more than one term, therefore, during the second term most of these students had a different instructor. These students were scattered throughout the classes of the 16 instructors teaching the second term of the three-term sequence. Some had failed the first term's work and repeated it, hence they were in classes of one of the three instructors teaching the first term of the course. These instructors were requested to rate the students on their ability to think scientifically by

using the rating sheet described above. In all, a total of 19 instructors were involved in the rating of the students. The two instructors that taught the students during the first term were responsible for most of the ratings. Each second term instructor rated a few students only.

In order to use these ratings in statistical computation, composite ratings were calculated for each student. A very superior rating was allotted 5 points, a superior rating 4 points, an average rating 3 points, an inferior rating 2 points, and a very inferior rating 1 point. Since each student was rated on two abilities by two judges, a maximum of 20 points and a minimum of 4 points was the range of possible scores on the composite rating.

An expectancy chart,<sup>5</sup> which reveals the expected performance of persons receiving various test scores, was one of the methods used to describe the validity of the test in terms of the rating of students by their instructors. A double entry table was constructed with scores on the test as one axis and scores on the ratings as the other axis. Because there were very few rated by both raters as either very superior or very inferior, the expectancy chart was constructed on the basis of superior, average, and inferior ratings. Rating scores from 10 through 14 were considered average, scores below 10 were considered inferior, and

---

<sup>5</sup> Adkins, op. cit., pp. 163-164.

scores above 14 were considered superior.

The expectancy chart can be treated statistically by means of the chi-square test.<sup>6</sup> The hypothesis to be tested was that the scores made by the students on Test IA were essentially unrelated to the ratings of the students by their instructors on their ability to think scientifically. The expectancy charts, Tables XXXII and XXXIII, show the observed numbers of persons in each category and, in parenthesis, the numbers which would be expected in each of the categories if there were no relationship between the scores on Test IA and the ratings.

TABLE XXXII

EXPECTANCY CHART SHOWING THE COMPARISON  
OF SCORES ON THE TEST IA PRE-TEST AND RATINGS

Scores	Ratings			Totals
	Superior	Average	Inferior	
75 - 100	14* (3.5)**	8 (12.6)	0 (5.9)	22
50 - 74	9 (10.6)	51 (37.8)	6 (17.5)	66
24 - 49	0 (8.9)	23 (31.5)	32 (14.6)	55
Totals	23	82	38	143

Degrees of Freedom - 4

Chi-square - 83.179

For these data chi-square is significant  
at the 1 percent level at 13.277

\* Observed number

\*\* Expected number

<sup>6</sup> Henry E. Garrett, Statistics in Psychology and Education. New York: Longmans, Green & Company. 1947. pp. 252-253.

The expectancy chart for the scores of the students in the pre-test group is presented as Table XXXII. The chi-square for these data was 83.179. Since a chi-square of 13.277 is required to make the results significant at the one percent level, it is evident that the hypothesis that there was no relationship between the test score and the rating must be rejected. On the contrary, there was a highly significant relationship between the scores on Test IA and the ratings of the students by the judges.

TABLE XXXIII

EXPECTANCY CHART SHOWING THE COMPARISON  
OF SCORES ON THE TEST IA POST-TEST AND RATINGS

<u>Scores</u>	<u>Ratings</u>			<u>Totals</u>
	<u>Superior</u>	<u>Average</u>	<u>Inferior</u>	
80 - 104	21* (6.9)**	19 (24.2)	0 (9.0)	40
55 - 79	2 (12.4)	56 (43.5)	14 (16.1)	72
30 - 54	0 (3.8)	6 (13.3)	16 (4.9)	22
Totals	23	81	30	134

Degrees of Freedom - 4

Chi-square - 84.471

For these data chi-square is significant  
at the 1 percent level at 13.277

\* Observed number

\*\* Expected number

The expectancy chart for the scores made by the students in the post-test group is presented as Table XXXIII.

The discrepancy in numbers in Table XXXII and Table XXXIII is due to the fact that a number of students were absent during the period when the test was given the second time. It is of some interest to note that the inferior group had a large number of absences. The chi-square for these data was 84.47 supporting the inference that there was a highly significant relationship between the scores on Test IA and the rating of the students by the judges. These findings give evidence that the test was valid providing the ratings of the judges were valid.

A comparison of the means of these three groups on the two administrations of the test is of interest. These are presented in Table XXXIV.

TABLE XXXIV

MEAN GAINS OF STUDENTS RATED AS SUPERIOR,  
INFERIOR AND AVERAGE ON TEST IA

Ratings	Group						
	Pre-test			Post-test			Gains
	No.	Mean	S. D.	No.	Mean	S. D.	
Superior	22	77.00	10.10	22	92.00	7.07	15.00
Average	82	57.10	12.63	81	70.45	12.12	13.35
Inferior	38	39.39	8.18	30	55.17	12.15	16.22

The differences between the means and the critical ratios of these differences were calculated. Table XXXV gives evidence that the group rated as superior was superior.

on performance on the test to a highly significant degree and the performance of the group rated as inferior was poorer than the performance of the group rated as average to a highly significant degree.

TABLE XXXV

DIFFERENCES IN MEANS AND CRITICAL RATIOS OF DIFFERENCES BETWEEN STUDENTS RATED SUPERIOR AND STUDENTS RATED AVERAGE AND STUDENTS RATED AVERAGE AND STUDENTS RATED INFERIOR

Group	Superior Dif. in mean	- Average C.R.	Average Dif. in mean	- Inferior C.R.
Pre-test	19.90	10.59	21.55	12.75
Post-test	17.73	10.75	15.28	8.08

Table XXXIV is also of interest in that it gives evidence that the increase in scores discussed previously in this chapter was not restricted to any particular group; the means of all of the groups, superior, average, and inferior being higher after a term of Biological Science.

The final method used to indicate the validity of the test was the determination of validity coefficients. Coefficients of correlation were calculated between total scores on the rating scale and (1) scores made on the test prior to taking Biological Science and, (2) scores made on the same test after one term of Biological Science. These correlations were  $.77 \pm .04$  and  $.72 \pm .04$  respectively. Such correlations

give evidence that the test had a considerable degree of validity insofar (in all these comparisons) as one could assume that the judges' ratings were a valid measure of the ability to think scientifically.

## CHAPTER VII

### SUMMARY AND CONCLUSIONS

#### SUMMARY

1. The purpose of this study was to devise a valid test to measure some of the inductive aspects of the ability to think scientifically, in the area of biological science.

2. The educational objectives to be measured by the test were formulated from Keeslar's<sup>1</sup> list of elements of scientific thinking. These objectives were:

- I. The ability to sense a problem.
  - II. The ability to state a problem.
  - III. The ability to delimit a problem.
  - IV. The ability to recognize facts which are related to the problem.
  - V. The ability to formulate hypotheses.
  - VI. The ability to plan experiments to test hypotheses.
  - VII. The ability to carry out experiments.
  - VIII. The ability to interpret data.
  - IX. The ability to formulate generalizations based on data.
  - X. The ability to apply generalizations to new situations.
3. The objectives were defined in terms of desired

---

<sup>1</sup> Oreon Keeslar, "The elements of scientific method." Science Education, 29:273:278, December, 1945.



behaviors involved in scientific thinking. In all, 98 behaviors were recognized as attending the skills of scientific thinking.

4. Situations in which the student could be expected to display the behaviors defined were identified. The sources of such situations were popular and scientific journals, textbooks, and interviews with members of the Department of Biological Science of Michigan State College.

5. Techniques for obtaining evidence concerning the attainment of the educational objectives were developed. In some instances techniques used previously were utilized and, in other cases, new techniques were devised.

6. Nine tryout tests, consisting of a total of 637 items, were constructed. These nine tryout tests were intended to measure respectively:

Test A. Some Steps in Scientific Thinking.

Test B. The Delimitation of Problems.

Test C. Experimental Procedures.

Test D. Organization of Data.

Test E. Evaluation of Hypotheses.

Test F. Experimentation and the Interpretation  
of Data.

Test G. Drawing of Conclusions.

Test H. Interpretation of Data.

Test J. Generalizations and Assumptions.

7. The tryout tests were administered to 168 students

during the spring term of 1950. The means, standard deviations, and reliabilities were calculated for each of the tryout tests. The reliabilities, determined by the method of split-halves with correction by the Spearman-Brown formula, ranged from .59 to .93. The mean, standard deviation, and the reliability of the tryout tests considered as a single test were determined. The reliability determined by the Kuder-Richardson formula was  $.92 \pm .01$ .

8. Item validity and item difficulty were calculated for each item of the tryout tests. The scores on each of the tryout tests were used as the criteria for item analysis. The purpose of these determinations was to identify those items of the tryout tests which were sufficiently discriminating and of suitable difficulty to be included in a single test, The Ability to Think Scientifically.

9. In order to determine whether there was a sufficient overlapping in the tryout tests to justify the elimination of some of the types of items in the construction of the single test, The Ability to Think Scientifically, intercorrelations of all of the tryout tests were calculated. These intercorrelations ranged from .11 to .59. Intercorrelations corrected for attenuation ranged from .17 to .73.

10. Coefficients of determination were calculated to determine the degree of overlapping of the tests. The degree of overlapping among the tryout tests ranged from 3 percent

to 53 percent. These amounts of overlapping seemed to indicate that there was not sufficient duplication to justify the elimination of any of these types of items.

11. In order to determine whether any one of the tryout tests was sufficiently similar to the score on the battery of tests to justify its use instead of the score on the tryout battery, the scores on each of the tryout tests were correlated with the total score on the tryout test battery. These correlations ranged from .41 to .74 indicating that all of the test had some relationship to the criterion (total tryout test scores) but that no single test measured all of the abilities appraised by the battery.

12. Multiple correlations between the total tryout test scores and each combination of two of the individual tryout tests were correlated. These multiple correlations ranged from .54 to .87, showing some of the pairs of tests were fairly adequate measures of the abilities involved in scientific thinking, whereas other pairs were quite inadequate.

13. A multiple correlation between the total tryout test scores and seven of the nine tryout tests was calculated by the Wherry-Doolittle method. A multiple correlation of .977 was obtained. These data gave evidence that the abilities could be measured quite adequately by less tests than had been used in the tryout test battery.

14. Correlations between the scores on the tryout

tests and the scores on the quantitative portion of the American Council on Education Psychological Examination ranged from .17 to .38, while correlations between the scores on the tryout tests and the scores on the linguistic portion of the American Council on Education Psychological Examination ranged from .11 to .43. Correlations between the scores on the tryout tests and the scores on the American Council on Education Reading Test ranged from .10 to .41.

15. Test I, The Ability to Think Scientifically, a single test of 150 items, was constructed from items of the tryout tests. This test was administered to 500 students who had completed three terms of Biological Science at the end of the spring term of 1950, and to 240 students at the beginning of the fall term of 1950 who had had no Biological Science. The means, standard deviations, and reliabilities of this test were determined for both groups. The reliabilities of Test I for the two groups were .89 and .91 respectively.

16. Item validities and item difficulties were calculated for each item of this test, using the total score on the test as the criterion for the item analysis.

17. Test I proved too long to be completed in a single laboratory period of one hour and fifty minutes. Therefore, Test IA, The Ability to Think Scientifically, was constructed from Test I, by the deletion of twenty-five of

the poorer items as determined by item analysis. This test was administered at the beginning of the fall term to 330 students who had had no Biological Science and to 136 of these same students at the end of the fall term of 1950. The means, standard deviations, and reliabilities were determined for the entire group and for the part of the group who took the test again at the end of the term. The reliabilities were  $.91 \pm .01$  and  $.90 \pm .02$  respectively.

18. The curricular validity of the test was established by:

1. Designing the test items to measure the behaviors involved in scientific thinking.
2. Submission of the tryout tests to competent judges for criticism.
3. Using free responses of students as items wherever feasible.
4. Careful selection of materials utilized in the construction of the test items.

19. The statistical validity of the test was established by:

1. Comparison of scores made on the tests with scores made on tests of (a) intelligence, (b) reading ability, and (c) knowledge of facts.
2. Comparison of scores made by students having had no Biological Science with scores made by students having had Biological Science.
3. Comparison of scores made on the test with ratings of the students by their instructors on their ability to think scientifically.

20. The correlation between the scores on the total tryout test battery and scores on the American Council on Education Psychological Examination was .51. Scores on Test I and the Psychological Examination gave the same correlation. The correlation between scores on the tryout test battery and scores on the American Council on Education Reading Test was .49, whereas the correlation of Test I with the reading test was .43.

21. Since the tests of the ability to think scientifically and the intelligence test both involved reading ability, partial correlations, with reading ability partialled out, were calculated. The partial correlation of the tryout tests was .31 while the partial correlation for Test I was .33. In order to determine the degree to which reading ability and the ability to think scientifically were related partial correlations, with intelligence partialled out, were calculated. For the tryout tests this partial correlation was .28; for Test I the partial correlation was .18.

22. The correlation between scores on the total tryout test battery and the portion of the comprehensive examination used to measure overall achievement in basic Biological Science which tested knowledge of facts was .33.

23. Scores made by the 500 students who took Test I after three terms of Biological Science were compared with another group of 240 students who had had no Biological

Science. The difference between the means of these two groups was highly significant. Test IA was given as a pre-test to 136 students before taking Biological Science and as a post-test to these same students after completion of one term of Biological Science. The difference between the means for the pre-test and the post-test was also highly significant, giving some evidence that if the test was a valid measure of the ability to think scientifically, the ability could be improved as a result of instruction.

24. One hundred and forty-three students taught by the present investigator and one other instructor in Biological Science were rated by means of the rating scale presented in Appendix IV on their ability to think scientifically. These students were a part of the 330 students who were given Test IA at the beginning of the Fall term of 1950. As previously mentioned, 136 of these students were given Test IA as a post-test at the end of the Fall term of 1950. These students were also rated on their ability to think scientifically by the instructors who taught them during the Winter term of 1951.

25. The chi-square test revealed that there was a significant relationship between the scores made on Test IA, both as a pre-test and as a post-test, and the averaged ratings of the judges.

26. The difference between the means of the test

for those students rated as superior and the means for those students rated as average was highly significant. So also was the difference of the means of those rated as average and those rated as inferior.

27. The correlation between scores on the pre-test and the ratings of the judges was  $.77 \pm .04$ . Between scores on the post-test and the judges' ratings the correlation was  $.72 \pm .04$ .

### CONCLUSIONS

On the basis of these findings the conclusion may be drawn that the test, The Ability to Think Scientifically, was sufficiently reliable for individual use, and that the test had sufficient validity to be used as a measure of the ability to think scientifically.

The data presented here support the inferences drawn from findings of previous studies that there is a moderate positive relationship between the ability to think scientifically and (1) intelligence, (2) reading ability, and (3) knowledge of facts. The findings of this study also support the inference that the ability to think scientifically is subject to improvement when this is a specific objective of instruction.



## EDUCATIONAL IMPLICATIONS

Educational implications for Biological Science at Michigan State College. The test, The Ability to Think Scientifically, should be useful for the appraisal of the teaching of the scientific method in Biological Science at Michigan State College. The laboratory studies at Michigan State College have been written with the expressed objective of teaching the scientific method. However, no valid test had been available to appraise the laboratory studies now being taught and to appraise the value of studies which may be written in the future.

The test should also be useful for diagnostic purposes. The test might be administered as a pre-test, and students making low scores on this pre-test might profit from remedial instruction in this area.

A pre-test program, to determine which students should be allowed to take the comprehensive examination after the completion of one term instead of after the three-term sequence, might include this test as a measure of one of the objectives of the course.

The relative merit of various methods of teaching scientific thinking might be evaluated by the test. Some have claimed that a course without a lecture would implement this objective, while others have claimed that a lecture -

demonstration method would be as effective as any other. These and other methods might be appraised by use of this test designed to measure the ability to think scientifically.

The findings of a significant gain in scores after taking Biological Science may stimulate further educational research. An experiment should be carried out to determine what factors are responsible for the increase in scores.

Educational implications for science courses in general education. Since the ability to think scientifically is a stated objective of almost all science courses in the general education program the test, The Ability to Think Scientifically, might be useful in other courses in biology at other institutions or modified for use for courses in the physical sciences.

Those test items which present a new technique for evaluating the ability to think scientifically may stimulate further work in the development of tests to measure this ability. The test as designed is probably too difficult for use in the secondary school; however, some of the techniques may be useful to persons constructing tests for secondary school use.

The findings of improvement in the ability to think scientifically, although not in itself conclusive, since no control group was included in this study, tend to support the conclusion that scientific thinking can be taught. The

accumulating evidence for this conclusion has far-reaching educational implications, and encourages educators to make further efforts to implement this important objective.

Other educational implications. The test, The Ability to Think Scientifically, might have some value for prediction of success in the field of science, or the techniques presented here might be modified in the construction of such tests. The present need for detecting of future scientists might be in some measure met by portions of this test.

Some of the techniques used in this test might also be modified for the construction of tests of critical thinking in other areas, such as the social sciences.

#### PROBLEMS SUGGESTED BY THE STUDY

Since the purpose of this study was to construct a reliable and valid measure of the ability to think scientifically, the study presented more problems than it solved. It was not the primary purpose of this study to investigate educability in the ability to think scientifically, nor was it the purpose of this study to investigate the relationship of ability to think scientifically to other traits such as reading ability and intelligence. However, in the validation of the test, some data relating to the above mentioned problems were accumulated. These data suggest a number of

problems.

The very evident question which arises from this study is "Did instruction in scientific thinking cause the significant increase in scores on the test?" A controlled experiment should be conducted. One group should be taught by the method used in Biological Science at Michigan State College, a second group should be taught the same subject matter by traditional methods and a third group should receive no science training. Such an experiment might indicate whether the laboratory program in Biological Science with the teaching of the scientific method as its major objective is more effective in evoking changes in behavior than traditional methods. It should also throw light on the question of whether ability to think scientifically is a by-product of the teaching of science. In addition, it should show whether improvement in the ability to think scientifically is merely a growth or maturation process.

Another problem arises from the finding of a moderate correlation between the test, The Ability to Think Scientifically, and intelligence. The problem suggested is "What factors of intelligence are related to the ability to think scientifically." In order to answer this question Thurstone's test of Primary Mental Abilities and the test, The Ability to Think Scientifically, could be given to a group of students. Factor analysis might reveal the loadings of various factors

in the test.

The finding of a correlation of .33 between knowledge of facts and ability to think scientifically indicates that some relationship exists between knowledge of facts and the ability to think scientifically. Since the factual test used in this correlation was not over the same subject matter as the test itself, this may not reflect the true relationship, which might be higher than reported in this study. In the construction of the test reported in this study it was assumed that students knew some general biological facts and vocabulary, such as the terms vitamin and bacteria. This assumption may not have been valid; therefore, a test should be devised to measure knowledge of the facts and vocabulary which were assumed to be general information. This information test should be administered just prior to the administration of the test, The Ability to Think Scientifically. A correlation between the two tests might reveal a more valid relationship between knowledge and the ability to think scientifically.

Test IA was administered to 136 students as a pre-test and as a post-test after one term of Biological Science. A few students made lower scores on the post-test than on the pre-test but most of the students made gains. These gains ranged from one to 41 points. This was not unusual; a test, retest situation almost always shows a similar trend.

However, the question may be asked, "Why do a few students fail to make any gain, while a few others make gains of almost one hundred percent of this original score?" Although the variation might be due to chance, the problem seems to be worth investigating. Several hypotheses are suggested. It might be that those students who participate actively in the laboratory program made large gains while those who do not participate actively in the laboratory program make small gains. This hypothesis could be tested by individual case study. A few students could be observed carefully by instructors and ratings of their acceptance of the objective of the laboratory program correlated with gains on the test.

Another hypothesis is that there may be a relationship between gains on the test, and gain in knowledge of biological facts. This hypothesis could be tested by giving pre-tests and post-tests. The test of the ability to think scientifically, and a test of knowledge of biological facts taught in the course could be used. Gains on the two tests could be correlated.

As discussed in Chapter IV, the test of ability to do scientific thinking was limited to a measurement of the critical aspects of scientific thinking. This limitation of the problem suggested a field of investigation which is of interest. What is the relationship between critical

thinking and creative thinking? What part does critical thinking play in creative thinking? How can creative thinking be measured reliably and validly?

Problems of technique in test construction are also suggested by this study. These are (1) to devise a valid and reliable test to be administered in fifty minutes, (2) to devise a test centered about a single problem, and (3) to devise several forms of such a test. A shorter test would be desirable if it were to be used as a pre-test and as a post-test each term, since it would not necessitate the use of two entire laboratory periods. A test revolving around a single problem would aid in the integration of the materials, while several forms of the test would reduce the possibility of memory playing a part in observed increase in scores.

## LITERATURE CITED

## BOOKS

- Adkins, Dorothy C. Construction and Analysis of Achievement Tests. Washington: U. S. Government Printing Office. 1947. Pp. 292.
- Aikin, Wilford M. The Story of the Eight-Year Study. New York: Harper and Brothers. 1942. Pp. 157.
- American Council on Education, Executive Committee of the Cooperative Study in General Education. Cooperation in General Education. Washington: American Council on Education. 1947. Pp. 220.
- Baten, William D. Elementary Mathematical Statistics. New York: John Wiley and Sons. 1938. Pp. 338.
- Bond, Austin D. M. An Experiment in the Teaching of Genetics with Special Reference to the Objectives of General Education. Contributions to Education, No. 797. New York: Bureau of Publications, Teachers College, Columbia University. 1940. Pp. 99.
- Buros, Oscar K. The Nineteen Forty Mental Measurement Yearbook. Highland Park, New Jersey: The Mental Measurement Yearbooks. Pp. 674.
- Curtis, Francis D. Some Values Derived from an Extensive Reading of General Science. Contributions to Education, No. 163. New York: Bureau of Publications, Teachers College, Columbia University. 1924. Pp. 142.
- Daily, Benjamin W. The Ability of High School Pupils to Select Essential Data in Solving Problems. Contributions to Education, No. 190. New York: Bureau of Publications, Teachers College, Columbia University. 1925. Pp. 103.
- Davis, Frederick B. Item-Analysis Data. Cambridge: Graduate School of Education, Harvard University. 1946. Pp. 42.
- Dewey, John. How We Think. Boston: D. C. Heath and Company. 1909. Pp. 244.



Educational Policies Commission, Education for All American Youth. Washington: National Education Association. 1944. Pp. 421.

Flanagan, John C. Critical Requirements for Research Personnel. Pittsburg: American Institute for Research. 1949. Pp. 66.

Gans, Roma. A Study of Critical Reading Comprehension in the Intermediate Grades. Contributions to Education, No. 811. New York: Bureau of Publications, Teachers College, Columbia University. 1940. Pp. 135.

Garrett, Henry E. Statistics in Psychology and Education. New York: Longmans, Green and Company. 1947. Pp. 487.

General Education in the American College. Thirty-eighth Yearbook of the National Society for the Study of Education, Part II, Pp. 380. Bloomington, Illinois: Public School Publishing Company, 1939.

Glaser, Edward M. An Experiment in the Development of Critical Thinking. Contribution to Education, No. 843. New York: Bureau of Publications, Teachers College, Columbia University. 1941. Pp. 212.

Gray, William S., editor. Recent Trends in American College Education. Chicago: University of Chicago Press. 1931. Pp. 249.

Harvard University. General Education in a Free Society. Cambridge: Harvard University Press. 1945. Pp. 257.

Hawkes, Herbert E., E. F. Lindquist, and C. R. Mann, The Construction and Use of Achievement Examinations. Cambridge: Houghton Mifflin Company. 1936. Pp. 497.

Judd, Charles H. Education as Cultivation of the Higher Mental Processes. New York: The Macmillan Company. 1936. Pp. 201.

McCall, William A. Measurement. New York: The Macmillan Company. 1939. Pp. 535.

McNemar, Quinn. Psychological Statistics. New York: John Wiley and Sons. 1949. Pp. 364.

Muskingum College. A College Looks at its Program. Columbus: The Spahr and Glen Company. 1937. Pp. 306.

- Noll, Victor H. The Habit of Scientific Thinking. A Handbook for Teachers. New York: Bureau of Publications, Teachers College, Columbia University. 1935. Pp. 27.
- Noll, Victor H. The Teaching of Science in the Elementary and Secondary Schools. New York: Longmans, Green and Company. 1939. Pp. 238.
- Program for Teaching Science. Thirty-first Yearbook for the National Society for the Study of Education, Part I, Pp. 364. Bloomington, Illinois: Public School Publishing Company. 1932.
- President's Commission on Higher Education. Higher Education for American Democracy. Volume I. Establishing the Goals. New York: Harper and Brothers. 1947. Pp. 103.
- Progressive Education Association. Science in General Education. New York: D. Appleton-Century Company. 1938. Pp. 591.
- Remmers, Hermann H. and N. L. Gage. Educational Measurement and Evaluation. New York: Harper and Brothers. 1943. Pp. 560.
- Science Education in American Schools. Forty-sixth Yearbook of the National Society for the Study of Education, Part I, Pp. 298. Chicago: The University of Chicago Press. 1947.
- Smith, Eugene R., Ralph W. Tyler and the Evaluation Staff. Appraising and Recording Student Progress. New York: Harper and Brothers. 1942. Pp. 550.
- Spafford, Ivol, editor. Building a Curriculum for General Education. Minneapolis: The University of Minnesota Press. 1943. Pp. 409.
- Stroud, James B. Psychology in Education. New York: Longmans, Green and Company. 1946. Pp. 664.
- Tyler, Ralph W. Constructing Achievement Tests. Columbus, Ohio: Ohio State University. 1934. Pp. 110.
- Tyler, Ralph W. Service Studies in Higher Education. Columbus, Ohio: Ohio State University. 1932. Pp. 283.
- Watson, Goodwin B. The Measurement of Fairmindedness. Contributions to Education, No. 176. New York: Bureau of Publications, Teachers College, Columbia University. 1925. Pp. 97.

## MONOGRAPHS AND BULLETINS

- Conrad, Herbert S. Characteristics and Use of Item-Analysis Data. American Psychological Association, Psychological Monographs: General and Applied. No. 295. 1948. p. 15.
- National Education Association. Reorganization of Science in Secondary Schools. U. S. Bureau of Education Bulletin, 1920, No. 26, Washington: Government Printing Office. Pp. 62.

## PERIODICAL LITERATURE

- Alpern, Morris L. "The ability to test hypotheses." Science Education, 30:220-229, October, 1946.
- Barnard, J. Darrell. "The lecture-demonstration versus the problem-solving method of teaching a college science course." Science Education, 26:121-132, October, 1942.
- Arnold, Dwight. "Testing ability to use data in the fifth and sixth grades." Educational Research Bulletin, 17:255-259, December, 1937.
- Blair, Glenn M. and Max R. Goodson. "Development of scientific thought in general science." School Review, 47:696-700, November, 1939.
- Beauchamp, Wilber L. "Teaching scientific method." School Science and Mathematics, 34:508-510, May, 1934.
- Billings, Marion L. "Problem solving in different fields of endeavor." American Journal of Psychology, 46:259-272, April, 1934.
- Bingham, Eldred N. "A direct approach to the teaching of the scientific method." Science Education, 33:241-249, April, 1949.
- Burke, Paul J. "Testing for critical thinking in physics." American Journal of Physics, 17:527-532, December, 1949.
- Committee on Research in Secondary-School Science. "Problems related to the teaching of problem-solving that need to be investigated." Science Education, 34:180-184, April, 1950.

- Crowell, Victor L. Jr. "The scientific method." School Science and Mathematics, 37:525-531, May, 1937.
- Curtis, Francis D. "Teaching scientific methods." School Science and Mathematics, 37:816-819, November, 1934.
- Davis, Ira C. "Is this the scientific method?" School Science and Mathematics, 34:83-86, January, 1934.
- Dewey, John. "Method in science teaching." Science Education, 29:119-123, April, 1945.
- Downing, Elliot R. "The elements and safeguards of scientific thinking." Scientific Monthly, 26:231-243, March, 1928.
- Downing, Elliot R. "Some results of a test on scientific thinking." Science Education, 20:121-128, October, 1936.
- Downing, Elliot R. "Teaching scientific method." School Science and Mathematics, 34:400-405, March, 1934.
- Edwards, Thomas B. "Measurement of some aspects of critical thinking." Journal of Experimental Education, 18:263-279, March, 1950.
- Engelhart, Max D. and Hugh B. Lewis, "An attempt to measure scientific thinking." Educational and Psychological Measurement, 1:289-294, Third Quarter, 1941.
- Flanagan, John C. "General Considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution." Journal of Educational Psychology, 30:674-680, December, 1939.
- Frutchey, Fred P., Ralph W. Tyler and B. Clifford Hendricks. "Measuring the ability to interpret experimental data." Journal of Chemical Education, 13:62-64, February, 1936.
- Grant, Charlotte L. and Elsa M. Meder. "Some evaluation instruments for biology students." Science Education, 28:106-110, March, 1944.
- Grener, Norma and Louis E. Raths, "Thinking in third grade." Educational Research Bulletin, 24:38-42, February, 1945.

- Grim, Paul R. "Interpretation of data and reading ability in social studies." Educational Research Bulletin, 19:372-374, September, 1940.
- Hart, E. H. "Measuring critical thinking in a science course." California Journal of Secondary Education, 14:334-338, October, 1939.
- Hered, William and Herbert A. Thelen. "The high-school chemistry test of the Armed Forces Institute." Journal of Chemical Education, 21:507-515, October, 1944.
- Herring, John P. "Measurement of some abilities in scientific thinking." Journal of Educational Psychology, 9:535-558, December, 1918.
- Higgins, Conwell D. "The educability of adolescents in inductive ability." Science Education, 29:82-85, March, 1945.
- Howell, William S. "The effect of high school debating on critical thinking." Speech Monographs, 10:96-102, Annual, 1943.
- Johnson, Alma. "An experimental study in the analysis and measurement of reflective thinking." Speech Monographs, 10:83-96, Annual, 1943.
- Keeslar, Oreon. "A survey of research studies dealing with the elements of scientific method as objectives of investigation in science." Science Education, 29:212-216, October, 1945.
- Keeslar, Oreon. "The elements of scientific method." Science Education, 29:273-278, December, 1945.
- Kelley, Truman L. "The selection of upper and lower groups for the validation of test items." Journal of Educational Psychology, 30:17-24, January, 1939.
- Le Sourd, Homer W. "Teaching scientific method." School Science and Mathematics, 34:234-235, March, 1934.
- Mallison, George G. "The implication of recent research in the teaching of science at the secondary school level." Journal of Educational Research, 43:321-342, January, 1950.

- Neuhof, Mark. "Integrated interpretation of data tests." Science Education, 26:21-26, January, 1942.
- Noll, Victor H. "Teaching the habits of scientific thinking." Teachers College Record, 35:202-212, December, 1933.
- Raths, Louis E., "A thinking test." Educational Research Bulletin, 23:72-75, March, 1944.
- Read, John G. "A non-verbal test of the ability to use the science method as a pattern of thinking." Science Education, 33:361-366, December, 1949.
- Sinclair, James H. and Ruth S. Tolman. "An attempt to study the effect of scientific training upon prejudice and illogicality of thought." Journal of Educational Psychology, 24:362-370, May, 1933.
- Smith, Victor C. "A study of the degree of relationship existing between ability to recall and two measures of ability to reason." Science Education, 30:86-90, March, 1946.
- Strauss, Sam. "Some results of the test of scientific thinking." Science Education, 16:89-93, December, 1931.
- Teller, James D. "Improving ability to interpret educational data." Educational Research Bulletin, 19:363-371, September, 1940.
- Teichman, Louis. "The ability of science students to make conclusions." Science Education, 28:268-279, December, 1944.
- Ter Keurst, Arthur J. and Robert E. Bugbee. "A test on scientific method." Journal of Educational Research, 36:489-501, March, 1943.
- Tyler, Ralph. "Measuring the results of college instruction." Educational Research Bulletin, 11:253-260, May, 1932.
- Ullsvik, Bjarne R. "An attempt to measure critical judgment." School Science and Mathematics, 49:445-452, June, 1949.
- Wood, Ben D. and F. S. Beers. "Knowledge versus thinking." Teachers College Record, 37:487-499, March, 1936.

Weller, Florence. "Attitudes and skills in elementary science." Science Education, 17:90-97, April, 1933.

Zyve, D. L. "A test of scientific aptitude." Journal of Educational Psychology, 18:525-546, November, 1927.

## TESTS

Love, Kenneth G. "Scientific Attitude - Thinking." Every Pupil Test. Columbus, Ohio: The State Department of Education. April, 1937.

## UNPUBLISHED MATERIALS

Bedell, Ralph C. "The Relationship Between the Ability to Infer in Specific Learning Situations." Unpublished Doctor's thesis, Department of Education, University of Missouri. 1934. Pp. 54.

Dunning, Gordon M. "The Construction and Validation of a Test to Measure Certain Aspects of Scientific Thinking in the Area of First Year College Physics." Unpublished Doctor's thesis, Department of Education, Syracuse University. 1948. Pp. 108.

Edwards, Thomas B. "Measurement of Some Aspects of Critical Thinking." Unpublished Doctor's thesis, Department of Education, University of California. 1949. Pp. 200.

Fleming, Maurice C. "An Analytical Study of Certain Outcomes of a Course for Orientation in Biological Science at College Level." Unpublished Doctor's thesis, Department of Education, New York University. 1942. Pp. 324.

Furst, Edward J. "Changes in Organization of Various Abilities and Skills after Two Years of General Education at the Secondary-School Level." Unpublished Doctor's thesis, Department of Education, University of Chicago. 1948. Pp. 249.

Higgins, Conwell D. "Educability of Adolescents in Inductive Ability." Unpublished Doctor's thesis, Department of Education, New York University. 1942. Pp. 206.

- Hoff, Alfred G. "A Test for Scientific Attitude."  
Unpublished Master's thesis, Department of Education,  
University of Iowa. 1930. Pp. 156.
- Thelen, Herbert A. "An Appraisal of Two Methods for  
Teaching Scientific Method in General Chemistry."  
Unpublished Doctor's thesis, Department of Education,  
University of Chicago. 1944. Pp. 370.
- Weisman, Leah L. "Some Factors Related to the Ability  
to Interpret Data in Biological Science." Unpublished  
Doctor's thesis, Department of Education, University  
of Chicago. 1946. Pp. 176.



## APPENDIX I

## TEST A

## SOME STEPS IN SCIENTIFIC THINKING

This test is designed to measure your ability to differentiate phases of thinking. These steps include major problems or perplexities, possible solutions to problems, observations which are not results of experimentation but rather preliminary observations, results of experimentation, and conclusions.

Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space on the answer sheet.

Key

1. A major problem (either stated or implied).
2. Hypothesis (possible solution to problem).
3. Results of experimentation.
4. Observations (not experimental).
5. Conclusion (probable solution to problem).

Ever since the days of Hippocrates one of medicine's big mysteries has been (1) the bodily process that transforms disease into death. With a special type of equipment which makes blood vessels transparent and three dimensional under a microscope, one investigator began examining the blood of healthy animals. The (2) blood cells of the healthy animals are separate and move rapidly. One day while observing the blood of a monkey dying of malaria, this researcher saw that the (3) blood was flowing slowly. Its consistency changed before his eyes. The blood (4) cells began to clump together in sluggish masses. The investigator realized that this (5) altered blood might be a major cause in the animal's illness. If the blood changes could

occur in malaria they might occur

Abbreviated Key

in other diseases as well - perhaps

1. A major problem

all diseases. The investigator

2. Hypothesis

3. Results

studied the circulation in other

4. Observations

5. Conclusions

diseased animals and found this clumping of blood (which

he called "sludged") in every diseased animal and those

suffering from severe injury or disease. (6) What makes

the red cells stick together? It was seen that (7) during

disease and injury the body deposits a sticky substance on

blood cells, causing the blood cells to stick together and

clog the circulation. If the process continues unchecked

death occurs. Other workers had seen sludged blood before

but its significance had been missed. This researcher thinks

that red cells (8) clumping may account for many cases of

mental illness, since he has found (9) in a few psychiatric

patients plugs in the brain indicating that there has been

sludging at one time. He also suggests that aging and

senility may (10) be accounted for by accumulated damage

from injury and illness. The discovery that sludge is a

critical factor in many diseases may prove to be one of the

great accomplishments of medicine. It opens up new ways of

fighting disease. (11) To find drugs to break up the sludge

and (12) to discover why sludge forms are two approaches

which are being followed.

The following key is to be used for the succeeding paragraph. Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space in the answer sheet.

Key

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Result of experimentation.
4. Initial observation (not experimental).
5. Conclusion (probable solution of problem).

(13) How does a homing pigeon navigate over territory it has never seen before? (14) Do air currents stimulate the pigeon in some way? (15) Are the pigeons equipped with some sort of magnetic compasses; that is, are they sensitive to the earth's magnetism? Yeagley tested the latter by fastening small magnets to the wings of well-trained pigeons. (16) Most of these birds never got home. (17) Others, carrying equal wing weights of non-magnetic copper, made the home roost without trouble, (18) indicating that the earth's magnetism is a factor in pigeon navigation. But the pigeons magnetic compass could not, by itself, bring him back to his roost; because many places on the earth's surface have identical magnetic conditions. Leagley endeavored (19) to determine the other guiding factor. (20) It might be the sun or stars, but pigeons navigate under clouds. While looking at a map which had lines representing the intensity of the earth's magnetism, he noted that the lines were crossed at varying angles by the parallels of latitude. (21) If pigeons are sensitive to some factor connected with the lines of latitude, they would have all they need to find their way

home. The next step was (22) to find some physical force, something the pigeons might be able to detect, related to the lines of latitude.

# Abbreviated Key

1. A major problem
2. Hypothesis
3. Results
4. Observations
5. Conclusions

The effect of the earth's turning varies directly with latitude; objects near the equator are carried daily around the earth's circumference, moving at over 1,000 mi. per hr. Objects near the poles are carried around more slowly. The direction and variation of this circling can be recorded by various man-made instruments. (23) Why shouldn't the pigeons feel it, too? (24) If they could, they would have, along with their magnetic compass a satisfactory navigating instrument. Yeagley trained hundreds of pigeons to return to their home roosts at State College, Pa. Then he took them to a part of Nebraska where the lines representing the earth's magnetism cross the parallels of latitude at the same angle as at State College. He released the pigeons to the east of this spot. (25) The pigeons all flew west. Yeagley believes that (26) pigeons are guided by both the earth's magnitude and by its turning. (27) Just where the birds keep their instruments is still unknown; but Yeagley found that (28) birds have a mysterious organ in their eyes, at the end of the optic nerve. (29) This organ may contain the nerve fibers that pick up vibrations of magnetism and the even more delicate sense that measure the earth's turning.

The following key is to be used for the succeeding paragraph. Certain parts of the paragraph are underlined, and each underlined portion is an item of the test. Choose the proper response from the key and blacken the appropriate space on the answer sheet.

Key

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Results of experimentation.
4. Initial observation (not experimental).
5. Conclusion (probable solution of problem).

(30) The residents of Deaf Smith County, Texas, are amazingly free of tooth decay. (31) The vegetables grown in this county are also unusual in that they attain a huge size. (32) Tooth decay has always puzzled scientists. (33) Could there be a relationship between the eating of these vegetables and the prevention of tooth decay? (34) Could the milk in this area be better for teeth? (35) Was the water in some way responsible for both the freedom from decay and the size of the vegetables? In Bausite, Ark., dentists noted that (36) most of the residents had blemishes on their teeth. Analysis of the water showed (37) it contained fluorine. There was little doubt that (38) the fluorine was responsible for the blemishes. Dentists also noticed that the (39) children of the community had almost no cavities in their teeth. On the assumption that (40) tooth decay is related to the amount of fluorine in the water, fluorine was used in a weak solution to paint the teeth and gums of half of the children in a community where no fluorine is normally found in the water. (41) These children had 40% less cavities than the children not

receiving the treatment. Dental  
 researchers have continued (42)  
the search for the essential cause  
of decay. (43) Diet deficiencies

Abbreviated Key

1. A major problem
2. Hypothesis
3. Results
4. Observations
5. Conclusion

have always been considered to be a major factor in tooth  
decay, but investigators found that (44) 124 patients  
suffering from diseases caused by dietary deficiencies had  
only one-third as many cavities as well-fed people. But  
 why? It has been found (45) that a certain germ called  
Lactobacillus acidophilis is found in the saliva of persons  
with many cavities, while it is practically absent from the  
mouths of those without cavities. One experimenter fed a  
 group of people with large numbers of these germs in the  
 mouths a six-week diet low in sugars and starches. He  
 found that (46) there were very few of the germs in the  
mouths of these people from six months to two years after  
the discontinuation of the treatment. He believes (47)  
that a sugarless diet may encourage the growth of other  
germs which fight the Lactobacillus acidophilis. That  
 (48) the prime cause of tooth decay is this Lactobacillus  
 is supported by the fact that flourine is very potent in  
 reducing the number of them in the mouth.

The following key is to be used in the paragraphs below. Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space on the answer sheet.

### Key

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Results of experimentation.
4. Observations (not experimental).
5. Conclusion (probable solution to problem).

The (49) sense least understood is the sense of smell. It has been generally believed that (50) the nose identified odors by chemical analysis. Some scientists suggested (51) that it is more likely that smelling is a measuring of infra-red (Heat) rays absorbed by odorous vapors. It has long been known that many gases absorb certain wave lengths of infra-red. (52) Chemists shoot infra-red rays through vapor and note what wave lengths are absorbed. (53) Why shouldn't the human nose do the same? In a study of substances which have odors and those which do not have odors they found that (54) all of those waves between  $7\frac{1}{2}$  to 14 microns long which do have odors can absorb infra-red whereas those without odors do not absorb these infra-red wave-lengths. Since the human body at normal temperature radiates heat waves chiefly at the  $7\frac{1}{2}$  to 14 band it may be that the ability to absorb heat waves is what makes vapors smellable. (55) But how does the nose do the smelling? The smell receptors in the upper nose lie across air passages. These researchers suggest that (56) when pure air is passing through the



nostrils the cells give no signal;

Abbreviated Key

they get rid of their heat at a  
standard rate. (57) But when an  
odorous vapor is present in the

1. A major problem.
2. Hypothesis
3. Results
4. Observations
5. Conclusions

air it absorbs certain wave lengths of heat from the cells.

(58) The cells feel the change and the stimulus produces a  
sensation of smell. To confirm this, these scientists,  
 studied cockroaches which have their smell receptors on their  
 antennae (hence outside the body). Cockroaches were known  
 to be attracted by oil of cloves. They put cockroaches in  
 a gas tight box with a window made of a material which was  
 transparent to infra-red. (50) The cockroaches responded  
just as strongly as if the window were not there, they  
swarmed toward the window. Then a window of glass, which  
 does not allow infra-red to go through it was put in as a  
 barrier. (60) The cockroaches showed no more interest in  
the window than if the oil of cloves were not there.

Next the researchers tried bees. (61) The bees  
crawled all over the heat-transparent window with sweet  
smelling honey vapor behind it, whereas (62) they ignored  
the window which did not allow the heat waves to pass  
through. Both (63) cockroaches and bees could smell vapors  
at a distance from their antennae. This may explain how  
 (64) some creatures, such as male moths seeking females,  
seem able to detect odors from considerable distance.

The following key is to be used in the following paragraph. Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space on the answer sheet.

Key

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Results of experimentation.
4. Observations (not experimental).
5. Conclusion (probable solution to problem).

High blood pressure and hardening of the arteries now afflict twice as many people as they did in 1900. (65) To find some cause for this increase in deaths much research has been conducted. (66) These conditions seem to run in families. (67) Are the conditions inherited? (68) Apparently diet is a factor in the production of the conditions because hardening of the arteries has been produced in rats by feeding them a diet high in cholesterol, a fat substance found in foods. Some scientists believe that although (69) people have a wonderful system to cope with emergencies, the unrelenting stress of civilized life is too much for it. The primary causes of these degenerative disorders, says one worker, (70) are overwork, fear and exposure to the elements. Any one of these may cause the pituitary gland at the base of the brain to pour more of its secretion into the blood stream. The pituitary secretion then stimulates the adrenal glands located above each kidney. (71) Normally the adrenal secretion causes a temporary rise in blood pressure during these times of crisis. This worker believes that if the crisis persists hardening of the arteries results. (72)

This experimenter noticed that

Abbreviated Key

people who had died after lives of

1. A major problem

tension had abnormally large

2. Hypothesis

3. Results

adrenal glands. He then subjected

4. Observations

5. Conclusion

animals to tensions to see if they developed similar

degenerative diseases. (73) He found that they did.

Although this work does not give a complete answer to

what causes degenerative diseases, (74) it does give

evidence that physically man is not quite adapted to

the civilization he has built.

TABLE XXXVI  
ITEM ANALYSIS DATA FOR TEST A

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*100.0	86.7	.55			
	**100.0	83.3	.50	33	91	78
2	82.2	35.5	.48			
	77.8	19.4	.60	42	48	49
3	82.2	71.1	.15			
	77.8	63.9	.17	10	70	61
4	86.7	62.6	.31			
	83.3	52.8	.35	22	69	60
5	93.3	77.8	.29			
	91.7	72.2	.32	20	82	69
6	97.8	88.9	.30			
	97.2	86.1	.32	20	91	79
7	71.1	37.7	.34			
	63.9	22.2	.43	28	43	46
8	91.1	62.2	.39			
	88.9	52.8	.43	28	70	61
9	82.2	46.7	.39			
	77.8	33.3	.46	30	55	53
10	91.1	53.3	.47			
	88.9	41.7	.54	36	64	58
11	77.8	51.1	.30			
	72.2	38.9	.34	21	55	53
12	75.6	60.0	.18			
	69.4	50.0	.22	13	59	55
13	93.3	82.2	.24			
	91.7	77.8	.24	15	85	72

\* Method of Flanagan  
\*\* Method of Davis

TABLE XXXVI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	86.7 83.3	40.0 25.0	.51 .58	40	55	53
15	88.9 86.1	40.0 25.0	.54 .61	43	55	53
16	95.6 94.4	77.8 72.2	.37 .39	25	83	70
17	95.6 94.4	73.3 66.7	.42 .45	29	80	68
18	73.3 66.7	37.7 22.2	.37 .46	30	44	47
19	100.0 100.0	80.0 75.0	.60 .58	40	86	73
20	95.6 94.4	75.6 69.4	.38 .41	27	82	69
21	77.8 72.2	53.3 41.7	.27 .31	19	57	54
22	95.6 94.4	73.3 66.7	.43 .45	29	80	68
23	51.1 38.9	8.9 0	.50 .67	49	19	32
24	64.4 55.6	44.4 30.6	.22 .26	16	42	46
25	77.8 72.2	62.6 52.8	.18 .20	12	61	56
26	62.6 52.8	24.4 5.6	.39 .58	40	28	38
27	93.3 91.7	62.6 52.8	.45 .50	33	71	62
28	84.4 80.6	51.1 38.9	.38 .43	28	59	55

TABLE XXXVI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
29	88.9 86.1	51.1 38.9	.46 .51	34	61	56
30	100.0 100.0	82.2 77.8	.60 .55	37	88	75
31	100.0 100.0	91.1 88.9	.50 .41	27	94	83
32	97.8 97.2	68.9 61.1	.54 .60	42	79	67
33	86.7 83.3	40.0 25.0	.51 .59	41	53	52
34	91.1 88.9	35.5 19.4	.60 .69	51	53	52
35	91.1 88.9	40.0 25.0	.57 .65	47	57	54
36	97.8 97.2	77.8 72.2	.47 .50	33	85	72
37	33.3 16.7	8.9 0.0	.36 .50	33	09	21
38	73.3 66.7	42.2 27.8	.32 .38	24	46	48
39	93.3 91.7	75.6 69.4	.32 .35	22	80	68
40	82.2 77.8	60.0 50.0	.27 .31	19	64	58
41	95.6 94.4	86.7 83.3	.25 .26	16	89	76
42	100.0 100.0	84.4 80.6	.55 .52	35	89	76
43	66.7 58.3	44.4 30.6	.23 .29	18	44	47

TABLE XXXVI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
44	75.6 69.4	33.3 16.7	.43 .54	36	42	46
45	68.9 61.1	17.8 0.0	.52 .77	61	30	39
46	95.6 94.4	86.7 83.3	.25 .23	14	89	76
47	86.7 83.3	73.3 66.7	.20 .22	13	74	64
48	84.4 80.6	73.3 66.7	.16 .17	10	73	63
49	73.3 66.7	42.2 27.8	.33 .38	24	46	48
50	86.7 83.3	40.0 25.0	.52 .59	41	53	52
51	100.0 100.0	71.1 63.9	.65 .66	48	82	69
52	20.0 0.0	13.6 0.0	.12 .00	0	0	0
53	57.8 47.2	20.0 0.0	.40 .71	54	24	35
54	22.2 2.8	11.1 0.0	.18 .20	12	02	8
55	100.0 100.0	93.3 91.7	.45 .38	23	95	85
56	91.1 88.9	51.1 38.9	.49 .55	37	63	57
57	68.9 61.1	42.2 27.8	.27 .34	21	44	47
58	77.8 72.2	44.4 30.6	.36 .40	26	51	51
59	95.6 94.4	77.8 72.2	.37 .39	25	83	70

TABLE XXXVI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
60	97.8	68.9	.54			
	97.2	61.1	.59	41	79	67
61	95.6	57.8	.55			
	94.4	47.2	.59	41	70	61
62	95.6	51.1	.60			
	94.4	38.9	.64	46	66	59
63	40.0	17.8	.27			
	25.0	0.0	.58	40	13	26
64	48.8	35.5	.14			
	36.1	19.4	.20	12	28	38
65	88.9	75.6	.22			
	86.1	69.4	.23	14	77	66
66	86.7	77.8	.16			
	83.3	72.2	.15	9	77	66
67	48.8	15.6	.38			
	36.1	0.0	.66	48	18	31
68	75.6	46.7	.32			
	69.4	33.3	.36	23	51	51
69	73.3	33.3	.41			
	66.6	16.7	.51	34	42	53
70	75.6	53.3	.25			
	69.4	41.7	.29	18	55	53
71	64.4	15.6	.52			
	55.6	0.0	.75	58	28	38
72	88.9	40.0	.54			
	86.1	25.0	.61	43	55	53
73	100.0	88.9	.50			
	100.0	86.1	.46	30	93	81
74	*75.6	48.8	.29			
	**69.4	36.1	.34	21	53	52

\* Method of Flanagan

\*\* Method of Davis



## TEST B

## THE DELIMITATION OF PROBLEMS

This test is designed to test your ability to delimit a problem. A problem is presented. This is followed by a series of questions. Rate the questions according to the following key.

Key

1. This question must be answered in order to solve the problem.
2. This question if answered might be useful in the solution of the problem.
3. The answer to this question, though related to the problem, would not help in the solution of the problem.
4. This question is completely unrelated to the problem.
5. This question if answered in the affirmative is a basic assumption of the problem.

PROBLEM: What causes colds?

## QUESTIONS:

1. Do all people have colds?
2. If one stays in bed with a cold does he get over the cold more rapidly?
3. Does one person "catch" a cold from another person who has a cold?
4. Why do some people have many colds and other people have few colds?
5. Is it possible to determine the cause of a cold?
6. Is there a germ present in persons with colds and absent from persons without colds?
7. Does aspirin help to cure a cold?
8. Can some germ be isolated which, when injected, will cause a cold?

- |  |                                |
|--|--------------------------------|
| 9. Do colds have a cause?  | <u>Abbreviated Key</u>         |
| 10. Does getting one's feet wet cause a cold?                      | 1. Must be answered            |
|  | 2. Might be useful             |
| 11. Does becoming chilled after being overheated cause a cold?     | 3. Related, but would not help |
|  | 4. Unrelated                   |
|  | 5. A basic assumption          |
| 12. Why are colds more prevalent in the winter than in the summer? |                                |
| 13. Do other animals get colds?                                    |                                |
| 14. Are people who are tired more susceptible to colds?            |                                |
| 15. Are there people who do not have colds but who are "carriers"? |                                |
| 16. How can colds be prevented?                                    |                                |

The thymus gland is located in the chest cavity just above the heart. This gland is largest during the growing period and becomes progressively smaller after maturity.

PROBLEM: What is the function of the thymus gland?

QUESTIONS:

17. Does lack of activity of the gland cause it to become smaller?
18. What causes the gland to stop functioning?
19. Is the gland inactive after maturity?
20. Does the removal of the gland before maturity cause an animal to become mature earlier?
21. Does the gland have a function?
22. Can any substance be extracted from the gland which when injected into another animal cause growth?
23. Why does the gland decrease in size after maturity?
24. Does the removal of the gland from young animals stunt their growth?
25. If the gland is removed will the animal mature?

- |     |   |  |
|-----|---|--|
| 26. | Do animals or people ever have disorders of this gland? | <u>Abbreviated Key</u><br>1. Must be answered<br>2. Might be useful<br>3. Related, but would not help<br>4. Unrelated<br>5. A basic assumption |
| 27. | Does the gland ever completely disappear?               |  |
| 28. | Can the function of the gland be determined?            |  |
| 29. | What are the effects of the removal of the gland?       |  |
| 30. | What causes the gland to grow smaller?                  |  |

A plant appeared which was different from its parents. The parent plants are essentially alike.

PROBLEM: What caused the plant to be different from its parents?

#### QUESTIONS:

31. Were the parent plants from pure lines; that is, were all of the known ancestors of both parents like the parents?
32. How does the plant differ from its parents?
33. Was the soil in which this plant was grown the same as the soil in which the parents were grown?
34. Why did this plant differ from its parents?
35. Was the difference due to the effects of the environment?
36. When did the change occur?
37. Will this plant produce seeds which when planted grow into plants like it?
38. Is it possible to determine what caused the change?
39. Is the change due to some change in the hereditary make-up of the plant, i.e., was it due to mutation?
40. What kind of a plant is it?
41. Do all plants produce offspring which are different from the parents?
42. Under what circumstances did the change occur?

43. Under what conditions did the plant develop? Abbreviated Key
44. Why would any plant be like its parents? 1. Must be answered  
2. Might be useful  
3. Related, but would not help
45. Was there any reason why the plant was different from its parents? 4. Unrelated  
5. A basic assumption
46. Were any of the ancestors like this plant?
47. Was the difference due to difference in the amount of sunlight the plant had?
48. How does this plant benefit man?

Bacterial cultures are frequently grown on the surface of a gelatin-like substance poured into a flat, covered dish. Occasionally these bacterial cultures become contaminated with molds. One scientist observed that bacteria did not grow in the vicinity of a certain green mold.

PROBLEM: What caused the bacteria-free zone around the mold?

QUESTIONS:

49. What kind of a mold was it?
50. Is there a relationship between the presence of the mold and the absence of bacteria?
51. Does the mold use the bacteria as food?
52. What kind of bacteria were they?
53. Is mold of any use to man?
54. Do the bacteria cause any disease?
55. Is some substance produced by the mold which kills the bacteria?
56. Why is the mold green?
57. Had the cultures of bacteria been properly prepared?
58. Was there any reason for the bacteria not being in the vicinity of the mold?

- |  | <u>Abbreviated Key</u>         |
|--|--------------------------------|
| 59. Do all molds cause bacteria-free zones around them?                    | 1. Must be answered            |
| 60. Where did the molds come from?   | 2. Might be useful             |
|  | 3. Related, but would not help |
| 61. Do bacteria produce any substance which kill the mold?                 | 4. Unrelated                   |
|  | 5. A basic assumption          |
| 62. Does the mold harm the growth of the cultures?                         |                                |
| 63. What substances cause bacteria-free zones?                             |                                |
| 64. Under what conditions were the cultures kept?                          |                                |
| 65. What is the green mold composed of?                                    |                                |
| 66. Did the green mold kill the bacteria or did it only stop their growth? |                                |
| 67. Would the green mold injure the cells of animals?                      |                                |

TABLE XXXVII  
ITEM ANALYSIS DATA FOR TEST B

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*53.3 **41.7	33.3 16.7	.22 .29	18	28	38
2	24.4 5.6	13.3 0.0	.17 .29	18	04	11
3	33.3 16.7	17.8 0.0	.20 .50	33	08	21
4	53.3 41.7	40.0 25.0	.14 .20	12	33	41
5	77.8 72.2	26.7 8.3	.52 .66	48	40	45
6	75.6 69.4	57.8 47.2	.20 .23	14	59	55
7	42.2 27.8	17.8 0.0	.29 .61	43	15	28
8	48.8 36.1	26.7 8.3	.24 .39	25	22	34
9	73.3 66.7	24.4 5.6	.49 .66	48	35	42
10	80.0 75.0	53.3 41.7	.31 .35	22	59	55
11	80.0 75.0	44.4 30.6	.38 .45	29	53	52
12	68.9 61.1	48.8 36.1	.22 .26	16	48	49
13	73.3 66.7	26.7 8.3	.47 .63	45	37	43

\* Method of Flanagan  
\*\* Method of Davis

TABLE XXXVII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	84.4	44.4	.45			
	80.6	30.6	.51	34	55	53
15	48.8	28.9	.21			
	36.1	11.1	.35	22	22	34
19	46.7	26.7	.22			
	33.3	8.3	.38	24	21	33
21	82.2	20.0	.62			
	77.8	0.0	.83	72	38	44
22	68.9	51.1	.19			
	61.1	38.9	.22	13	48	49
25	46.7	13.3	.40			
	33.3	0.0	.64	46	17	30
26	40.0	24.4	.18			
	25.0	5.6	.35	22	15	28
27	44.4	24.4	.22			
	30.6	5.6	.40	26	18	31
28	75.6	31.1	.45			
	69.4	13.9	.57	39	42	46
29	64.4	60.0	.05			
	55.6	50.0	.07	4	53	52
30	57.8	28.9	.30			
	47.2	11.1	.45	29	28	38
31	91.1	88.9	.04			
	88.9	86.1	.07	4	88	75
33	62.6	44.4	.19			
	52.8	30.6	.23	14	40	45
35	44.4	31.1	.15			
	30.6	13.9	.23	14	22	34
37	37.7	13.3	.32			
	22.2	0.0	.55	37	12	25

TABLE XXXVII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
38	77.8 72.2	24.4 5.6	.54 .70	53	38	44
39	71.1 63.9	60.0 50.0	.12 .15	9	57	54
40	46.7 33.3	37.7 22.2	.10 .14	8	28	38
42	55.6 44.4	35.5 19.4	.21 .29	18	31	40
43	60.0 50.0	31.1 13.9	.30 .41	27	31	40
44	26.7 8.3	8.9 0.0	.28 .35	22	05	14
45	55.6 44.4	17.8 0.0	.42 .70	52	22	34
46	51.1 38.9	20.0 0.0	.34 .67	49	19	32
47	75.6 69.4	55.6 44.4	.22 .26	16	57	54
51	68.9 61.1	33.3 16.7	.36 .47	31	38	44
53	86.7 83.3	80.0 75.0	.10 .12	7	79	67
54	80.0 75.0	60.0 50.0	.24 .26	16	61	56
56	51.1 38.9	40.0 25.0	.12 .14	8	31	40
57	35.5 19.4	24.4 5.6	.13 .29	18	12	25
58	71.1 63.9	20.0 0.0	.52 .78	63	31	40



TABLE XXXVII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
60	28.9	11.1	.27			
	11.1	0.0	.40	26	06	17
61	35.5	8.9	.37			
	19.4	0.0	.52	35	10	23
62	44.4	22.2	.25			
	30.6	2.8	.51	34	16	29
63	28.9	13.3	.23			
	11.1	0.0	.40	26	06	17
64	64.4	42.2	.23			
	55.6	27.8	.29	18	42	46
65	55.6	26.7	.30			
	44.4	8.3	.47	31	25	36
67	*68.9	48.8	.21			
	**61.1	36.1	.26	16	48	49

\* Method of Flanagan

\*\* Method of Davis

## TEST C

## EXPERIMENTAL PROCEDURES

This test is designed to measure your ability to recognize faulty experimental procedures and to test your ability to select the best of a series of experiments. In each case a problem and a possible solution to the problem (an hypothesis) are presented. In each case the experiments were designed by students to test the hypotheses. Judge each experiment according to the following key.

Key

1. This experiment is satisfactory.
2. This experiment is unsatisfactory because it lacks a control or comparison.
3. This experiment is unsatisfactory because the control or comparison is faulty.
4. This experiment is unsatisfactory because it is unrelated to the hypothesis.
5. None of the above - the experiment or situation is unsatisfactory for reasons other than those listed in 2, 3, and 4.

PROBLEM: What are some of the requirements for the sprouting of seeds?

## HYPOTHESIS:

Oxygen is a requirement for the sprouting of seeds.

1. Plant one seed in a container where oxygen is available and place another seed in a container where all oxygen has been removed. Keep all other conditions the same.
2. Put some seeds in soil in a flask from which all the oxygen has been removed. Put an airtight stopper in the flask to keep out all air. Then put some seeds of the same type in soil in a flask that is open and gets the oxygen from the air. See which sprouts or if both sprout. Keep moisture, temperature and amount of light, etc., the same in each flask.
3. If a seed lacked oxygen under a controlled experiment the seed would not function properly and would soon die.
4. Put two groups of seeds side by side, in the ground, only put a jar over one group to keep the oxygen away from them. Keep all other conditions the same for each group.

5. Take two groups of seeds each appropriately labeled and put one group in a compartment with the average amount of oxygen in normal conditions and an excess amount of oxygen in another sealed compartment. Keep all other conditions, such as light, moisture, etc., the same for each.
 

<u>Abbreviated Key</u>
1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to hypothesis
5. None of the above
6. Take two packages of seeds. Allow oxygen to be in contact with one package but keep the other package of seeds protected from all oxygen. Observe which sprouts.
7. Place growing plants in an air tight container. Pump out the oxygen. Place other growing plants in containers with oxygen. Keep temperature, light, etc., the same for each.
8. Plant seeds in a container with glass covering it so that no oxygen can enter and see if they sprout. Keep temperature, light and moisture normal.
9. Two groups of bean seeds might be set up. One in an air-tight container, absolutely free from oxygen. The other group could be allowed free circulation of air. After a specified length of time, the specimens could be examined and the need of oxygen for sprouting determined.
10. Set up two seed beds in which the moisture, temperature, amount of light, and all other factors are the same, except that the experimental seed bed has a very restricted supply of oxygen, while the control seed bed has a normal supply of oxygen.

PROBLEM: To determine the effects of a deficiency of Vitamin Y.

HYPOTHESIS: Vitamin Y affects the rate of growth of animals.

1. Get 40 young monkeys. Keep all vitamins from 20 of them, and feed the other 20 a normal supply of vitamins. Observe the weights and height of these monkeys for a year.
2. Take 60 young rabbits, divide them into three groups of 20 each. Feed the first group of 20 a normal diet of foods. Feed the second group a diet which contains

much Vitamin Y; feed the third group a diet completely devoid of Vitamin Y. Keep an accurate record of weights and length of the rabbits for 6 months.

# Abbreviated Key

- |   |                            |
|---|----------------------------|
|   | 1. Satisfactory            |
|   | 2. Lacks control           |
|   | 3. Control faulty          |
|   | 4. Unrelated to hypothesis |
| 3. Use two groups of young animals with all the conditions affecting the rate of growth of animals held constant and in one group supply Vitamin Y or omit Vitamin Y and observe the results in growth. | 5. None of the above       |
4. Take three normal young white rats. One is fed a well-balanced diet. Another is deprived of Vitamin Y only. The third is given an excess of Vitamin Y only. Make sure that all other conditions are kept the same.
  5. Find some animals that are naturally without an adequate supply of Vitamin Y. Try and find out why. From this you should be able to find out if Vitamin Y affects the rate of growth of animals.
  6. Take different kinds of young animals and to one kind feed a diet deficient in Vitamin Y, and to the other kind a diet rich in Vitamin Y. Measure and weigh the animals weekly.
  7. Give groups of animals identical diets for at least 2 weeks except for the omission of Vitamin Y from the diet of one group. Make sure all other factors - size, age, living conditions, etc., are the same for both. Make careful observations on weights of the animals.
  8. Start with 100 normal young animals. Make the diet of 50 of them deficient in Vitamin Y. Observe the differences between the two groups in rate of growth.
  9. Give Vitamin Y to a group of adults all about the same age. To another group of the same age give no Vitamin Y. Make certain that the diet and other living conditions of the two groups is the same in all other respects. Continue the experiment for a year. Keep weekly records of weight.
  10. Give Vitamin Y to a group of children who are living under favorable conditions (favorable to growth) and see whether Vitamin Y affects the growth of this group.

PROBLEM: A minute insect (aphid) is suspected of spreading a virus disease of roses. How would you determine whether this is true?

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to hypothesis
5. None of the above

HYPOTHESIS: The aphid spreads a virus disease of roses.

1. Put the insect among other kinds of plants other than roses. Leave another group of these plants free from contact with the aphids. Compare the results.
2. I would expose rats or guinea pigs to the roses to determine if the aphid is spreading a virus disease of roses. If the animals became ill then I would continue on to determine whether or not it was true.
3. Since aphids travel through the air, a plot of roses must be entirely protected from them, and another exposed to aphids which in turn have been exposed to roses afflicted with the virus disease. All must be under constant conditions of soil, atmosphere, etc.
4. Put some roses in a room; half which have the disease and half which do not. Put some of the aphids in the room. Observe and draw conclusions.
5. I would place 3 plants of roses in one room; one with a virus disease. In this room should also be the insects, aphid. Allow insects to go from infected plants to one of the other plants. These plants should be watched to see if the virus spreads.
6. Take sample rose with the virus disease. Obtain same kind of rose with no disease. Use microscope to aid in detection of the disease. Use some sort of spray. Note results.
7. Take 2 sets of the same kind of roses and expose one set of them to aphids. Keep the plants under the same conditions at all times and if the roses with the aphids contract the disease while the isolated ones do not, then the aphids are carriers.
8. Use rose plants which are known not to be diseased. In the same area place rose plants which are diseased but which have been treated to destroy the aphid. Note whether the disease still spreads after the aphids have been killed.

9. In order to determine whether the aphid spreads a virus disease in roses, a group of roses should be put in a hot house free from aphids to see whether they get such a virus disease.

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to hypothesis
5. None of the above

10. Select numerous roses free from the virus and from the same soil or area. Divide these and expose one group to aphids and isolate the other group. One would try to keep all environmental conditions for both groups alike with the exception of exposing the one group of roses to the insect.

PROBLEM: To find a prevention for disease X.

HYPOTHESIS: A newly developed vaccine will prevent the disease.

1. Use animals such as rats, rabbits and inject them with the vaccine. If it is successful then use on a human. The vaccine should be successful upon many people before the vaccine can be declared a preventive.
2. Use 500 people who have disease X and do nothing for them. Then take 500 other people who have disease X and give them the newly developed vaccine.
3. In an area where the disease is prevalent inject half of the population with the vaccine. Do not inject the other half. Make sure that the two groups are about the same in other ways. Compare the number of cases in the two groups.
4. With 4 guinea pigs, inject 2 of them with the vaccine, then place them in a contaminated place where they will be susceptible to the disease. Place 2 un-vaccinated in also. If the 2 un-vaccinated get the disease and the vaccinated do not - after several such experiments it is probable - if both get disease X, the vaccine is no good.
5. Put 2 animals into a region where disease X is prevalent. In one inject the vaccine and if the variable shows no signs of the disease the hypothesis is true.
6. Inject 20 mice with the new vaccine, leaving 20

untreated. Then inject all of the mice with disease X. If all of the mice get "X" vaccine is no good. If out of the 20 you injected with new vaccine none got the disease and the 20 control mice did, then you have a good vaccine.

### Abbreviated Key

- |  |                            |
|--|----------------------------|
|  | 1. Satisfactory            |
|  | 2. Lacks control           |
|  | 3. Control faulty          |
|  | 4. Unrelated to hypothesis |
|  | 5. None of the above       |
7. You have to make sure the vaccine cures the disease in animals first, as close to a natural condition in humans as possible. Then to try it on a human being and see if it reacts the same way. You cannot tell until you have tried it on a human.
  8. Inject a number of animals with disease X. With a similar needle inject the other half with the special vaccine. Note that everything must be the same except the vaccine. If the vaccine injected group does not get the disease and the others do it will substantiate the hypothesis.
  9. Take a diseased animal and expose him to a group of animals that have been innoculated with this new vaccine. Then take the same diseased animal and expose him to a group of healthy animals that have not been innoculated. (Rats preferably - but any type would suffice, as long as they are susceptible to the disease.)

PROBLEM: What are some of the requirements for the sprouting of seeds?

HYPOTHESIS: Seeds sprout within a certain temperature range.

1. Set up 7 seed beds in which the moisture, ventilation, and amount of light are the same. All factors are same, except one bed will be kept at 0°F., another at 20°, another at 40°, 60°, and 80°, 100°, and 120°F. Observe which sprout first, and which, if any, never do sprout. Try this same experiment with several different kinds of seeds.
2. Place seeds in various temperatures: warm, hot, cold, freezing, moderate. This will determine the range in which certain seeds will sprout. Different types of seeds may sprout in different temperature ranges. Keep all other conditions such as moisture, light, etc., the same.

3. One seed under temperature  
from 0-20°C.
- 2nd seed under temperature  
from 20-40°C.
- 3rd seed under temperature  
from 40-60°C.
- 4th seed under temperature  
from -20-0°C.

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to  
hypothesis
5. None of the above

Determine the temperature at which the seeds sprout the best.

4. Plant three seeds. Keep one above normal temperatures, the second below normal and the third at normal temperature. The seed that sprouts will tell which temperature range is the best.
5. Take about 10 sets of seeds of the same type planted in the same condition. Subject each set to a different temperature ranging from 0° to 100°C. Observe if seeds sprout at a certain temperature.
6. Put seeds in pots, and then put these pots in places where the temperature can be properly adjusted. Put one of these pots at every 10°, keeping all other conditions constant, some of these plants will sprout.
7. Put different seeds at varying degrees of temperature. See at which temperature they sprout.
8. Set up 2 conditions similar except one set of seeds planted would be placed where the temperature would be about 2°C, the other remain at about room temperature.
9. If the seed was placed in the earth at freezing temperatures it would not grow. I would say that the temperatures of 70°--100° would sprout the seed.
10. Take 2 groups of seed. Attempt to sprout seeds within this certain temperature. Attempt to sprout seeds under adverse temperature.



PROBLEM: To determine the cause of illness which appears when large numbers of people are being confined to a small space.

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to

HYPOTHESIS: Lack of oxygen causes the people to become ill.

- hypothesis
5. None of the above

1. Examine the ill people and trace back the illness to whether it is caused by lack of oxygen or what.
2. One might check the oxygen by placing a number of people in a confined place where there was a control amount. Other checks would have to be made also such as the purity of food, the purity of water and whether or not proper sanitation rules were followed.
3. Put 50 normal people into a small space under normal conditions. Put 50 normal people into a small space with a large forced supply of oxygen. Compare the two groups after a considerable time.
4. Take 50 monkeys or mice and put a group where the oxygen is low and put a group where the oxygen is kept higher. If lack of oxygen causes people to become ill it may make the monkeys or mice ill.
5. Have a person work and live normally in a room with insufficient oxygen. Another person work and live normally in a room with sufficient oxygen. Compare the effects.
6. Confine one group to a small space in which there is a limited supply of oxygen. Let the other group have unlimited supply of oxygen and a large space. Let their diets and other items be the same. If the cause of the illness is as stated the confined group will be ill from lack of oxygen.
7. Set two groups of people, one with plenty of oxygen and the other in a normal environment. Determine which group becomes ill.
8. Take 3 groups. Group 1 will be confined in small space, with the usual things. This is the control group. Group 2 will be confined in an equally small and crowded place, only they shall have excellent ventilation. Group 3 will be confined to spacious (relatively) quarters, and they shall not have good ventilation. Keep careful records and see what results suggest.

9. Put a lot of rabbits in a small space for a period of time. Put a few rabbits in the same amount of space. Observe the rabbits and draw conclusions.
10. First tests should be made on the air to see if there is a lack of oxygen. If there is a lack of oxygen and there is no other reason for the people being ill then the hypothesis would be true.
11. Observe the effects of a large number of people in a small room. Then add pure oxygen to the same room with the same people. If the illnesses were cured, it would be likely that the lack of oxygen was the cause.
12. Put different groups of people in different rooms. Give one group a greater amount of carbon dioxide than oxygen, the second group a normal amount and a third group a greater amount of oxygen than carbon dioxide and check for results.
13. Put one group of people in a room with an excessive amount of carbon dioxide and another group in a room with a normal amount of carbon dioxide. Keep the oxygen concentration the same in both rooms.
- Abbreviated Key
1. Satisfactory
  2. Lacks control
  3. Control faulty
  4. Unrelated to hypothesis
  5. None of the above

TABLE XXXVIII  
ITEM ANALYSIS DATA FOR TEST C

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*17.8 ** 0.0	0.0 0.0	.55 .00	0	0	0
2	77.8 72.2	66.7 58.3	.14 .15	9	64	58
3	71.1 63.9	17.8 0.0	.54 .78	63	31	40
4	66.7 58.3	48.8 36.1	.19 .22	13	46	48
5	33.3 16.7	17.8 0.0	.20 .50	33	08	21
6	64.4 55.6	15.6 0.0	.52 .75	58	28	38
7	71.1 63.9	28.9 11.1	.43 .56	38	37	43
8	97.8 97.2	64.4 55.6	.58 .61	43	77	66
9	77.8 72.2	68.9 61.1	.12 .14	8	66	59
10	57.8 47.2	51.1 38.9	.08 .08	5	43	46
11	33.3 16.7	17.8 0.0	.20 .50	33	08	21
12	93.3 91.7	91.1 88.9	.05 .07	4	90	77
13	60.0 50.0	48.8 36.1	.12 .15	9	42	46
14	4.4 0.0	11.1 0.0	-.18 .00	0	0	0

\* Method of Flanagan

\*\* Method of Davis

TABLE XXXVIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
15	44.4	22.2	.25			
	30.6	2.8	.51	34	16	29
16	73.3	31.1	.43			
	66.7	13.9	.55	37	40	45
17	2.2	6.7	-.16			
	0.0	0.0	.00	0	0	0
18	51.1	37.7	.14			
	38.9	22.2	.20	12	31	40
19	26.7	20.0	.13			
	8.3	0.0	.35	22	05	14
20	82.2	62.6	.24			
	77.8	52.8	.27	17	66	59
21	80.0	42.2	.40			
	75.0	27.8	.47	31	51	51
22	57.8	44.4	.14			
	47.2	30.6	.17	10	38	44
23	88.9	64.4	.34			
	86.1	55.6	.36	23	70	61
24	53.4	33.3	.21			
	41.7	16.7	.29	18	29	38
25	0.0	8.9	-.45			
	0.0	0.0	.00	0	0	0
26	68.9	33.3	.37			
	61.1	16.7	.47	31	38	44
27	15.1	2.2	.38			
	0.0	0.0	.00	0	00	0
28	64.4	40.0	.25			
	55.6	25.0	.32	20	40	45
29	86.7	44.4	.48			
	83.3	30.6	.54	36	57	54
30	13.3	8.9	.13			
	0.0	0.0	.00	0	0	0

TABLE XXXVIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
31	82.2 77.8	42.2 27.8	.43 .50	33	53	52
32	33.3 16.7	11.1 0.0	.32 .50	33	08	21
33	62.6 52.8	44.4 30.6	.19 .23	14	40	45
34	17.8 0.0	4.4 0.0	.31 .00	0	0	0
35	44.4 30.6	6.7 0.0	.50 .62	44	16	29
36	15.4 0.0	0.0 0.0	.55 .00	0	0	0
37	51.1 38.9	28.9 11.1	.23 .36	23	26	36
38	60.0 50.0	26.7 8.3	.34 .51	34	28	38
39	73.3 66.7	46.7 33.3	.27 .34	21	50	50
40	93.3 91.7	73.3 66.7	.35 .36	23	79	67
41	68.9 61.1	37.7 22.2	.32 .40	26	42	46
42	44.4 30.6	2.2 0.0	.64 .62	44	16	29
43	42.2 27.8	2.2 0.0	.62 .61	43	15	28
44	86.7 83.3	57.8 47.2	.35 .39	25	66	59
45	71.1 63.9	46.7 33.3	.25 .31	19	48	49
46	40.0 25.6	20.0 0.0	.24 .58	40	14	27

TABLE XXXVIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
47	48.8 36.1	40.0 25.0	.09 .14	8	30	39
48	53.3 41.7	37.7 22.2	.16 .23	14	31	40
49	33.3 16.7	13.3 0.0	.28 .50	33	08	21
50	35.5 19.4	33.3 16.7	.03 .05	3	17	30
51	71.1 63.9	22.2 2.8	.49 .71	54	33	40
52	17.8 0.0	22.2 2.8	-.03 -.17	-10	02	8
53	51.1 38.9	40.0 25.0	.12 .17	10	31	40
54	17.8 0.0	11.1 0.0	.11 .00	0	0	0
55	46.7 33.3	31.1 13.9	.17 .26	16	22	34
56	55.6 44.4	17.8 0.0	.43 .70	52	22	34
57	31.1 13.9	37.7 22.2	-.07 -.12	- 7	18	31
58	53.3 41.7	22.2 2.8	.34 .59	41	22	34
59	48.8 36.1	20.0 0.0	.32 .66	48	18	31
60	13.3 0.0	6.7 0.0	.10 .00	0	0	0
61	40.0 25.0	33.3 16.7	.08 .12	7	21	33
62	*84.4 **80.6	44.4 30.6	.44 .51	34	55	53


\* Method of Flanagan


\*\* Method of Davis


## TEST D


## ORGANIZATION OF DATA

This test is designed to test your ability to organize data. Select from the key below the curve which best fits the data. If none of the curves fit the data mark space five on your answer sheet.

- Key
1. 

2. 

3. 

4. 

5. none of the curves

1. The horizontal axis represents temperature. The vertical axis represents the amount of Substance A derived from Substance B.

<u>Temperature</u>	<u>Amount of Substance A</u>
10°C.	4 grams
25°C.	7 grams
35°C.	9 grams
60°C.	14 grams

2. The horizontal axis represents the amount of oxygen in the experimental gas mixtures. The vertical axis represents the amount of oxygen taken up by red cells in these experiments.





<u>Oxygen in gas mixtures</u>	<u>Oxygen taken up by red cells</u>
0	0
10	50
20	75
30	90
50	98

3. The horizontal axis represents the percent of carbon dioxide in gas mixtures breathed in; the vertical axis represents the percent increase in total amount of gas breathed per minute.

<u>Carbon dioxide percent</u>	<u>Percent increase</u>
0	0
1	10
2	25
3	50
5	100
7	200

4. The horizontal axis is the concentration of salt. (Sodium chloride). The vertical axis is the percent of red cells destroyed in these concentrations of salt.

Abbreviated Key

1.  3.   
 2.  4.  5. none

<u>Concentration of salt</u>	<u>Percent red cells destroyed</u>
.27	98
.36	75
.41	10
.50	1

5. The horizontal axis represents the amount of thyroprotein fed daily to cows. The vertical axis represents the percent increase in milk production.

<u>Thyropotein fed</u>	<u>Percent increase</u>
.15 grams	18
.20 grams	23
.24 grams	27
.30 grams	33

6. The horizontal axis represents age in years. The vertical axis is the percent increase in the weight of the brain from birth to twenty years of age.

<u>Age</u>	<u>Percent increase</u>
1 yr.	40
4	80
12	98





7. The horizontal axis represents the time in minutes to kill bacteria in a weak solution of silver nitrate. The vertical axis are the temperatures to which the bacteria in the silver nitrate solutions were subjected.

<u>Time in minutes</u>	<u>Temperature</u>
160	15°C.
80	20°C.
40	30°C.
0	45°C.



8. The horizontal axis represents age in years. The vertical axis is the percent increase in the weight of the ovaries and other female sex organs from birth to 20 years.

Abbreviated Key

1.  3.  5. none  
2.  4. 

Age

4  
10  
14  
18

Percent increase

8  
12  
20  
80

9. The horizontal axis represents time in seconds; the vertical axis represents the amount of heat developed in a single contraction of a single muscle fiber.

Time

.0  
.1  
.2  
.4  
.8

Heat

0.0  
10.0  
15.0  
14.0  
4.0

10. The horizontal axis represents the number of days since the memorization of certain nonsense syllables; the vertical axis is the percent of the nonsense syllables forgotten.

Time in days

1  
2  
6  
12

Percent forgotten

45  
60  
80  
84

11. The horizontal axis represents age of girls in years; the vertical axis is the strength index of these girls in pounds.

Age





1  
3  
8  
15  
20

Strength index

.5  
2.0  
5.0  
12.0  
12.5

12. The horizontal axis represents the successive number of trials in the learning of a puzzle. The vertical axis is the time in seconds of each trial.

Abbreviated Key

1.  3.   
 2.  4.  5. none

Trial

1st  
5th  
10th  
14th  
18th

Time in seconds

420  
419  
240  
60  
50

13. The horizontal axis represents the time in hours after the injection of sugar into the blood; the vertical axis is the amount of sugar in the blood.

Time after injection

1  
3  
6

Blood sugar

35  
12  
8

14. The horizontal axis is the time in minutes after pint jars of corn have been put in boiling water and kept boiling; the vertical axis is the temperature in the center of the pint jar.

Time

5  
10  
30  
60  
100

Temperature

20  
21  
55  
90  
99





15. The horizontal axis is age in years. The vertical axis is the metabolic rate of an individual expressed in calories per day.

Age

2  
5  
15  
25  
40

Calories

60  
40  
30  
25  
23

16. The horizontal axis represents time in days; the vertical axis is the number of yeast cells in millions (starting with 100 yeast cells).
- Abbreviated Key
1.  3.  5. none  
2.  4. 

<u>Time in days</u>	<u>Number of yeast cells in millions</u>
4	25
8	150
12	390
20	400

17. The horizontal axis is the temperature in Centigrade. The vertical axis represents the amount of enzyme activity of a certain type of bacteria in arbitrary units.

<u>Temperature</u>	<u>Enzyme activity</u>
10	0
30	1
50	2
70	3
90	2.5

18. The horizontal axis represents age in weeks; the vertical axis represents the weight of an animal, in kilograms.





<u>Age</u>	<u>Weight</u>
1	.05
3	.15
8	.80
12	1.6
16	2.4
25	2.8

19. The horizontal axis represents the external temperature; the vertical axis represents the amount of oxygen absorbed by a frog at the various temperatures.

<u>Temperature</u>	<u>Oxygen</u>
10	104 mg.
14	130 mg.
20	160 mg.
25	208 mg.

20. The horizontal axis represents the time in hours. The vertical axis the temperature inside a thermos bottle containing germinating pea seeds.

Abbreviated Key

1.  3.   
2.  4.  5. none

Time

0  
12  
24  
36

Temperature

20°C.  
24°C.  
30°C.  
32°C.

TABLE XXXIX

## ITEM ANALYSIS DATA FOR TEST D

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*75.6	42.2	.35			
	**69.4	27.8	.40	26	48	49
2	95.6	37.7	.67			
	94.4	22.2	.74	57	57	54
3	80.0	24.4	.56			
	75.0	5.6	.71	54	40	45
4	88.8	20.0	.68			
	86.1	0.0	.87	80	42	46
5	84.4	51.1	.38			
	80.6	38.9	.43	28	59	55
6	91.1	26.7	.66			
	88.9	8.3	.78	63	48	49
7	77.8	11.1	.67			
	72.2	0.0	.81	69	37	43
8	95.6	26.7	.73			
	94.4	8.3	.83	71	51	51
9	93.3	31.1	.67			
	91.7	13.9	.75	59	53	52
10	93.3	31.1	.67			
	91.7	13.9	.75	59	53	52
11	75.6	35.5	.41			
	69.4	19.4	.51	34	44	47
12	93.3	33.3	.65			
	91.7	16.7	.74	57	53	52
13	95.6	15.6	.80			
	94.4	0.0	.90	90	46	48

\* Method of Flanagan

\*\* Method of Davis

TABLE XXXIX (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	75.6	33.3	.43			
	69.4	16.7	.54	36	42	46
15	97.8	33.3	.75			
	97.2	16.7	.81	68	57	54
16	64.4	20.0	.46			
	55.6	0.0	.75	58	28	38
17	77.8	26.7	.52			
	72.2	8.3	.66	48	40	45
18	68.9	26.7	.43			
	61.1	8.3	.59	41	35	42
19	33.3	22.2	.14			
	16.8	2.8	.35	22	09	22
20	*42.2	22.2	.23			
	**27.8	2.8	.48	32	15	28

\* Method of Flanagan

\*\* Method of Davis

## TEST E

## EVALUATION OF HYPOTHESES

This test is designed to measure your understanding of the relation of facts to the solution of a problem. The over-all problem involved in this test is presented. This is followed by a series of possible solutions to the problem (hypotheses). After each hypothesis there are a number of items, all of which are true statements of fact. Determine how the statement is related to the hypothesis and mark each statement according to the key which follows the hypothesis.

## GENERAL PROBLEM:

What factors are involved in the transmission and development of Infantile Paralysis (Poliomyelitis)?

## HYPOTHESIS I.

In man the disease is contracted by direct contact with persons having the disease.

For items 1 through 11 mark space if the item offers:

1. direct evidence in support of the hypothesis.
2. indirect evidence in support of the hypothesis.
3. evidence which has no bearing on the hypothesis.
4. indirect evidence against the hypothesis.
5. direct evidence against the hypothesis.

1. Monkeys free from the disease almost never catch infantile paralysis from infected monkeys.
2. Most strains of infantile paralysis virus can be transferred from man only to monkeys and apes and not to other animals.
3. The virus has been isolated from the nasopharyngeal washings of humans and monkeys.
4. The curve of number of cases of the disease in a given area is the same shape as the curve for the fly population in that area, the infantile paralysis incidence curve lagging behind the fly population curve by about two weeks.
5. The virus has never been isolated from the blood.
6. The virus is not found in the nasal secretion, nor in the saliva.

7. The incubation period for infantile paralysis is from 4 to 21 days.
8. Most persons in contact with the diseased individual do not develop the disease.
9. The incidence of infantile paralysis is higher in rural districts than in the cities.
10. Cases of infantile paralysis have been found to follow the roads of communication of the population, that is, the disease spreads from populated areas along roads or rivers to other areas.
11. Even during epidemics cases are spotty, it is usually impossible to trace one case from another.
12. What is the status of hypothesis I ?
  1. It is true.
  2. It is probably true.
  3. It is false.
  4. It is probably false.
  5. The data are contradictory, hence its truth or falsity cannot be judged.

#### HYPOTHESIS II.

The disease is spread by the excrement (excreted material) of persons harboring the virus.

For items 13 through 23 mark space if the item offers:

1. direct evidence in support of the hypothesis.
  2. indirect evidence in support of the hypothesis.
  3. evidence which has no bearing on the hypothesis.
  4. indirect evidence against the hypothesis.
  5. direct evidence against the hypothesis.
13. The virus is always found in the stools of persons who have the disease.
  14. In the stools of persons not in contact with persons with the disease the virus is found in only one person in 100.
  15. During an epidemic non-paralytic cases outnumber paralytic cases ten to one.
  16. The curve of number of cases of the disease in a given area is the same shape as the curve for the fly population in the area, the infantile paralysis incidence curve lagging behind the fly population curve by about two weeks.



17. The incubation period for infantile paralysis is from 4 to 21 days.
18. Nine out of 14 adult contacts had virus in the stool, almost all child contacts have virus in the stools.
19. The virus has been isolated from streams carrying sewage.
20. Cases of the disease have been found to follow the roads of communication of the population, that is, the disease spreads from populated areas along roads or rivers to other areas.
21. The virus of the disease has been found in the stools and vomit of flies up to two days after eating an infected meal.
22. Even during epidemics cases are spotty.
23. It is usually impossible to trace one case from another.
24. What is the status of hypothesis II ?
  1. It is true.
  2. It is probably true.
  3. The data are contradictory, so the truth or falsity cannot be judged.
  4. The hypothesis is probably false.
  5. It is definitely false.

#### HYPOTHESIS III.

The olfactory nerve (nerve from nose to brain) is the route of entry of the virus.

For items 25 through 34 mark space if the item offers:

1. direct evidence in support of the hypothesis.
  2. indirect evidence in support of the hypothesis.
  3. evidence which has no bearing on the hypothesis.
  4. indirect evidence against the hypothesis.
  5. direct evidence against the hypothesis.
25. The virus has been isolated from nasopharyngeal washings of humans and monkeys.
  26. A plug of cotton, saturated with virus, placed in the nose of the monkey invariably causes the monkey to contract the disease. If the olfactory nerve is cut the monkey does not contract the disease when a plug saturated with the virus is placed in the nose.

27. If the nose of a monkey is sprayed with zinc sulphate the monkey (with virus plug inserted) does not contract the disease.
28. The virus is not found in the nasal secretion or in the saliva.
29. The virus has been isolated from the spinal cord of 71% of the cases autopsied, and from the olfactory nerve in 5% of the cases autopsied.
30. The virus has been found in the nasopharynx from several days before the onset of the disease until about 3 days after the onset of the disease.
31. Many doctors recommended the use of zinc sulphate nasal spray (administered only by the physician).
32. The virus is not affected by freezing.
33. Most strains of the virus can be transferred only to monkeys and apes.
34. The percentage of cases of infantile paralysis among persons receiving the nasal spray of zinc sulphate was the same as the percentage of cases in the total population.
35. What is the status of hypothesis III ?
  1. It is true.
  2. It is probably true.
  3. The data are contradictory, hence truth or falsity of the hypothesis cannot be judged.
  4. It is probably false.
  5. It is definitely false.

#### HYPOTHESIS IV.

The higher the degree of sanitation the greater are the chances of epidemic forms of the disease.

For items 36 through 45 mark space if the item offers:

1. direct evidence in support of the hypothesis.
  2. indirect evidence in support of the hypothesis.
  3. evidence which has no bearing on the hypothesis.
  4. indirect evidence against the hypothesis.
  5. direct evidence against the hypothesis.
36. Monkeys free of the disease almost never catch infantile paralysis from infected monkeys.

37. The virus has been isolated from streams carrying sewage.
38. In India epidemics seldom occur.
39. In India children under five are about the only ones affected.
40. During the war there was one epidemic among the European and American soldiers in India, the incidence among the soldiers was extremely high.
41. The percent of cases of infantile paralysis in whites is about four times that in colored people.
42. In the south (U.S.) there are three times as many cases under five years as over five years of age.
43. The percent of cases of infantile paralysis is higher in rural districts than in the cities.
44. In the north (U.S.) about 50% of the cases are over 5 years of age..
45. During an epidemic non-paralytic cases outnumber paralytic cases ten to one.
46. What is the status of hypothesis IV ?
  1. The hypothesis is true.
  2. It is probably true.
  3. The data are contradictory, hence the truth or falsity of the statement cannot be judged.
  4. It is probably false.
  5. It is definitely false.

#### HYPOTHESIS V:

Healthy persons having had contact with diseased individuals may carry the disease from one person to another.

For items 47 through 59 mark space if the item offers:

1. direct evidence in support of the hypothesis.
2. indirect evidence in support of the hypothesis.
3. evidence which has no bearing on the hypothesis.
4. indirect evidence against the hypothesis.
5. direct evidence against the hypothesis.

47. Monkeys free of the disease almost never catch infantile paralysis from infected monkeys.

48. During an epidemic non-paralytic cases outnumber paralytic cases ten to one.
49. It has been found that exertion prior to or at the time of infection increases the incidence of the disease.
50. Even during epidemics cases are spotty; it is usually impossible to trace one case from another.
51. The virus is always found in the stools of people who have the disease.
52. Most persons in contact with the diseased individual do not develop the disease.
53. Nine out of 14 adult contacts had virus in stools, almost all child contacts have virus in stools.
54. Up to two months after contact the virus is found in the stools of persons who contacted the victims, but who did not contract the disease.
55. In the stools of non-contacts the virus was found in only one person in 100.
56. Data on families each with one case of infantile paralysis in the family: 39% of other children in family from 1-4 years of age and 30% of other children in family 5-9 years of age had minor illnesses. Only 9% of children in other homes showed similar illnesses.
57. The percent of cases of infantile paralysis is higher in rural districts than in the cities.
58. Under twenty years of age the percent of cases in males is three times the percent of cases in females.
59. Flies were allowed to feed on contaminated food. The flies were then placed in contact with food which was fed to monkeys. The feces of the monkeys contained the virus.
60. What is the status of hypothesis V ?
  1. The hypothesis is true.
  2. It is probably true.
  3. The data are contradictory, so the truth or falsity cannot be judged.
  4. It is probably false.
  5. It is definitely false.

## HYPOTHESIS VI.

An immunity to the disease may be developed.

For items 63 through 71 mark space if the item offers:

1. direct evidence in support of the hypothesis.
2. indirect evidence in support of the hypothesis.
3. evidence which has no bearing on the hypothesis.
4. indirect evidence against the hypothesis.
5. direct evidence against the hypothesis.

61. Most strains of the infantile paralysis virus can be transferred only from man to monkeys and apes and not to other animals.
62. During an epidemic non-paralysis virus cases outnumber paralytic cases ten to one.
63. The incubation period of infantile paralysis is from 4 to 21 days.
64. Even during epidemics cases are spotty; it is usually impossible to trace one case from another.
65. Most persons in contact with the diseased individual do not develop the disease.
66. Up to two months after contact the virus is found in the stools of persons who contacted the victims, but who did not contract the disease.
67. In the stools of persons not in contact with persons with the disease the virus was found in only one person in 100.
68. Data on families each with one case of infantile paralysis in the family: 39% of the other children in family from 1-4 years of age and 30% of the other children in family from 5-9 years of age had minor illnesses. Only 9% of children in other homes showed similar illnesses.
69. Epidemics seldom occur in India and the disease is almost entirely among children under 5 years of age.
70. The percent of cases in whites is about four times the percent of cases in colored people.
71. Cases of infantile paralysis may continue into the winter, but an epidemic never arises anew during the winter.

72. What is the status of hypothesis VI ?
1. The hypothesis is true.
  2. It is probably true.
  3. The data is contradictory, so the truth or falsity cannot be judged.
  4. It is probably false.
  5. It is definitely false.
73. How many of the six hypotheses are acceptable?
1. 1
  2. 2
  3. 3
  4. 4
  5. 5
74. How many of the hypotheses are not acceptable?
1. 1
  2. 2
  3. 3
  4. 4
  5. 5

TABLE XXXX

## ITEM ANALYSIS DATA FOR TEST E

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*66.7	31.1	.38			
	**58.3	13.9	.47	31	35	42
2	57.8	40.0	.18			
	47.2	25.0	.24	15	35	42
3	40.0	31.1	.10			
	25.0	13.9	.17	10	19	32
4	53.3	17.0	.41			
	41.7	0.0	.69	51	21	33
5	77.8	44.4	.36			
	72.2	30.6	.40	26	51	51
6	53.3	22.2	.34			
	41.7	2.8	.60	42	22	34
7	93.3	66.7	.40			
	91.7	58.3	.45	29	74	64
8	91.1	84.4	.13			
	88.9	80.6	.15	9	83	70
9	82.2	40.0	.45			
	77.8	25.0	.55	37	53	52
10	57.8	44.4	.13			
	47.2	30.0	.18	11	38	44
11	55.6	37.7	.20			
	44.4	22.2	.24	15	33	41
12	48.8	22.2	.28			
	36.1	2.8	.56	38	19	32
13	71.1	71.1	.00			
	63.9	63.9	.00	0	64	58

\* Method of Flanagan

\*\* Method of Davis

TABLE XXXX (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	53.3 41.7	17.8 0.0	.38 .69	41	21	33
15	100.0 100.0	84.4 80.6	.50 .52	35	89	76
16	53.3 41.7	15.6 0.0	.41 .69	51	21	33
17	84.4 80.6	80.0 75.0	.05 .08	5	77	66
18	28.9 11.1	20.0 0.0	.12 .39	25	5	16
19	66.7 58.3	33.3 16.7	.36 .44	27	35	42
20	53.3 41.7	44.4 30.6	.08 .14	8	35	42
21	53.3 41.7	22.2 2.8	.34 .60	42	22	34
22	60.0 50.0	48.8 36.1	.12 .15	9	42	46
23	38.7 22.0	17.8 0.0	.27 .55	37	12	25
24	84.4 80.6	51.1 38.9	.38 .43	28	59	55
25	48.8 36.1	28.9 11.1	.21 .34	21	24	35
26	95.6 94.4	84.4 80.6	.25 .29	18	87	74
27	11.1 0.0	11.1 0.0	.00 .00	0	0	0
28	53.3 41.7	20.0 0.0	.36 .69	51	21	33
29	44.4 30.6	31.1 13.9	.15 .22	13	22	34



TABLE XXXX (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
30	57.8	33.3	.26			
	47.2	16.7	.38	24	31	40
31	46.7	24.4	.24			
	33.3	5.6	.41	27	19	32
32	97.8	84.4	.40			
	97.2	80.6	.41	27	89	76
33	91.1	75.6	.27			
	88.9	69.4	.27	17	79	67
34	55.6	37.7	.20			
	44.4	22.2	.24	15	33	41
35	62.6	24.4	.40			
	52.8	5.6	.58	40	28	38
36	75.6	44.4	.34			
	69.4	30.6	.39	25	50	50
37	17.8	17.8	.00			
	0.0	0.0	.00	0	0	0
38	55.6	17.8	.42			
	44.4	0.0	.70	52	22	34
39	31.1	20.0	.14			
	13.9	0.0	.46	30	8	20
40	46.7	24.4	.24			
	33.3	5.6	.43	28	19	32
41	62.6	28.9	.35			
	52.8	11.1	.47	31	31	40
42	26.7	26.7	.00			
	8.3	8.3	.00	0	8	20
43	46.7	13.3	.40			
	33.3	0.0	.65	47	17	30
44	15.6	13.3	.07			
	0.0	0.0	.00	0	0	0
45	100.0	77.8	.55			
	100.0	72.2	.61	43	85	72

TABLE XXXX (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
46	33.3 16.7	13.3 0.0	.30 .48	32	8	21
47	62.6 52.8	24.4 5.6	.40 .58	40	28	38
48	15.6 0.0	13.3 0.0	.07 .00	0	0	0
49	93.3 91.7	55.6 44.4	.49 .56	38	69	60
50	53.3 41.7	17.8 0.0	.49 .56	38	69	60
51	88.9 86.1	64.4 55.6	.33 .36	23	70	61
52	57.8 47.2	22.2 2.8	.38 .63	45	25	36
53	35.5 19.4	24.4 5.6	.14 .29	18	13	26
54	57.8 47.2	20.0 0.0	.40 .71	54	24	35
55	57.8 47.2	26.7 8.3	.32 .51	34	28	38
56	44.4 30.6	37.7 22.2	.06 .10	6	25	36
57	48.8 36.1	28.9 11.1	.21 .32	20	24	35
58	95.6 94.4	88.9 86.1	.20 .18	11	90	77
59	35.5 19.4	20.0 0.0	.18 .54	36	11	24
60	64.4 55.6	37.7 22.2	.28 .37	23	38	44

TABLE XXXX (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
61	46.7	40.0	.07			
	33.3	25.0	.10	6	28	38
62	60.0	46.3	.13			
	50.0	33.3	.17	10	42	46
63	95.6	88.9	.20			
	94.4	86.1	.18	11	90	77
64	37.7	35.5	.03			
	22.2	19.4	.02	2	21	33
65	68.9	24.4	.44			
	61.1	5.6	.64	46	34	41
66	48.8	35.5	.13			
	36.1	19.4	.17	10	28	38
67	66.7	42.2	.27			
	58.3	27.8	.31	19	42	46
68	48.8	35.5	.14			
	36.1	19.4	.20	12	28	38
69	51.1	33.3	.18			
	38.9	16.7	.27	17	27	37
70	46.7	24.4	.24			
	33.3	5.6	.43	28	20	32
71	73.3	46.7	.28			
	66.7	33.3	.34	21	50	50
72	71.1	37.7	.36			
	63.9	22.2	.43	28	42	46
73	42.2	13.3	.35			
	27.8	0.0	.61	43	15	28
74	*40.0	25.0	.17			
	**26.7	8.3	.31	19	17	30

\* Method of Flanagan

\*\* Method of Davis

## TEST F

## EXPERIMENTATION AND THE INTERPRETATION OF DATA

This test was designed to measure your ability to interpret data and to test your understanding of experimentation. In each case the numbers in the first column are the numbers which you will use as your answer. Thus the table presented becomes both the source of data and your key for the questions which follow it. In each case where a test tube number or group number is called for the one which gives positive evidence for the statement should be given. Below this the control or comparison is called for. This is the test tube or group number of the data which offers a comparison. For example:

1. Leaf in dark - no starch.
2. Leaf in light - starch.

Light is necessary for the production of starch. You would mark space 2 because this is the positive evidence, but it would be meaningless if it were not compared with the leaf in the dark. Therefore, the following item, "What is the control (comparison) for item 1?" would be marked space 1.

Items 1 through 15 refer to the data presented below. Some test tubes were set up and each contained 1 gram of fat. They were marked 1, 2, 3, 4, and 5. Mark each item according to the test tube number called for. Various substances were added to the tubes containing fat. All substances were dissolved in water before they were added to the fat. All test tubes were kept at 85° F. (Water boils at 212° F.) For test tube 5, Substance A was boiled and then allowed to cool before it was added to the fat.

Test Tube Number	Content of tube	Amt. of Substance B present after 24 hrs.
1	Fat plus Substance A	.1 gram
2	Fat plus Substance A plus Substance C	.5 gram
3	Fat plus Water	.0 gram
4	Fat plus Substance C	.0 gram
5	Fat plus Substance A (boiled)	.0 gram

1. Give the number of the test tube which acts as a control (comparison) for the entire experiment.
2. Give the number of the tube which gives evidence that fat does not break down spontaneously into Substance B in 24 hours.
3. Give the number of the tube used to show that a temperature of 85 degrees F. was not sufficient to cause fat to be broken down into Substance B.
4. Give the test tube number of the tube which gives evidence that Substance A is the active substance in the breakdown of fat to Substance B.
5. Give the test tube number of the tube which is the control (comparison) for item # 4.
6. Give the number of the tube which provides evidence that Substance C alone is ineffective in the breakdown of fats.
7. What is the control for item # 6?
8. Which test tube gives evidence that Substance C accelerates the rate of activity of Substance A?
9. Give the tube which is the control for item # 8.
10. Which tube gives evidence that Substance A is a substance whose properties can be destroyed?
11. Give the control for the tube in item # 10.
12. Which tube gives evidence that Substance C affects the fat in some way so that Substance A can more easily act upon it?
13. Which tube is the control for # 12?
14. Which tube gives evidence that Substance A is not a stable substance?
15. What is the control for item # 14?

Items 16 through 28 refer to the data presented below. Mark each item according to the group called for. Each group contained 100 persons fed on the diets indicated.

<u>Group</u>	<u>Diet</u>	<u>Cases of Beri Beri</u>
1	whole rice (i.e. rice with hulls)	none
2	polished rice (i.e. rice with hulls removed)	60%
3	polished rice plus Vitamin B <sub>1</sub>	none
4	polished rice plus Vitamin B <sub>2</sub>	60%
5	polished rice plus Vitamin B complex	none
16.	Give the number of the group which is the control (comparison) for the entire experiment.	
17.	Give the group which gives evidence that rice hulls contain a beri beri preventing substance.	
18.	Give the control for item 17.	
19.	Give the number of the group which provides evidence that Vitamin B is not a single entity.	
20.	Give the control for item 19.	
21.	Give the number of the group which indicate that rice hulls may contain Vitamin B.	
22.	Give the control for item 21.	
23.	Give the number of the group which provides evidence that rice hulls may contain Vitamin B <sub>1</sub> .	
24.	Give the number of the group which is the control for item 23.	
25.	Which group gives evidence that a differing of Vitamin B causes beri beri.	
26.	What is the control for item 25?	
27.	What group gives evidence that Vitamin B <sub>2</sub> is not the active factor in the prevention of beri beri?	
28.	What is the control for item 27?	

Items 29 through 39 refer to the data presented below. Mark each item according to the group number called for. When a person ascends to high altitudes his blood cell count increases after about 10 days. The following data were obtained from a study of altitude effects on rats. 760 mm. of mercury is atmospheric pressure at sea level. Air is composed of about 20% oxygen and 80% nitrogen.

<u>Group</u>	<u>Atmospheric pressure</u>	<u>% O<sub>2</sub></u>	<u>% N</u>	<u>Red cell count</u>
1	760	10	90	increased
2	380	20	80	increased
3	760	20	80	normal
4	760	40	60	decreased
5	380	40	60	normal

29. Give the number of the group which is the control for the entire experiment.
30. Give the number of the group that gives evidence that a decrease in atmospheric pressure causes an increase in red cell count at high altitude.
31. Which group is the control (comparison) for item 30?
32. Which group gives evidence that it is the decrease of oxygen pressure which is responsible for the increase in cell count at high altitudes?
33. Which of the groups is the best control for item 32?
34. Which of the groups gives evidence that a decrease in atmospheric pressure is not the cause of an increased red cell count at high altitudes?
35. What is the control for item 34?
36. Give the number of the group which gives evidence that a decrease in nitrogen pressure is not responsible for the increased red cell count at high altitudes.
37. Give the number of the group that is the control for item 36.
38. Which group gives evidence that an increase in oxygen pressure decreases the red cell count?
39. What is the control for item 38?

Items 40 through 57 refer to the data presented below. Mark each item according to the leaf number called for. Plant A normally stores starch in its leaves while plant B does not normally store starch in its leaves. The following experiments were performed in a dark room at 72 degrees F. Glucose (sugar) solutions were made with 20 grams of glucose per 100 cubic centimeters of water. Leaves of plant A taken from a plant that had been in the dark for 48 hours were floated in the 5 solutions listed below and left in the glucose solution for an hour.

<u>Leaf</u>	<u>Solution</u>	<u>Analysis of leaf after 4 hours</u>
1	Glucose	Starch in leaf
2	Water	No starch in leaf
3	Glucose plus juice from Plant B	No starch in leaf
4	Glucose plus juice from Plant C	No starch in leaf
5	Glucose plus boiled juice from Plant B	Small amount of starch in leaf

40. Give the number of the leaf which showed that starch does not develop spontaneously in the leaf in the dark.
41. This leaf indicates that a temperature of 72 degrees F. does not cause starch to form in the leaf.
42. Give the number of the leaf which is the control (comparison) for the entire experiment.
43. Give the number of the leaf which gives evidence that Plant A is capable of manufacturing starch from glucose.
44. Give the number of the leaf which is the control for item 43.
45. Give the number of the leaf which gives evidence that the juice of Plant B is capable of preventing the manufacture of starch from glucose.
46. What is the control for item 45?
47. Give the number of the leaf which gives evidence that Plant A is normally able to store starch in its leaves.
48. What is the control for item 47?



49. Give the number of the leaf which gives evidence that Plant C does not normally form starch in its leaves.
50. Give the leaf number of the control for item 49.
51. Which leaf shows that water does not cause the production of starch in the leaf?
52. Give the number of the leaf which gives evidence that the juices of Plant B contain a substance which inhibits the production of starch in its leaves.
53. Give the leaf which is the control for item 52.
54. This leaf gives evidence that the inhibitory substance is not a stable substance.
55. What is the control for item 54?
56. Give the number of the leaf which shows that boiling destroys the activity of the juice of Plant B.
57. Give the control for item 56.

Items 58 through 72 refer to the data presented below. Five test tubes, each containing a gram of protein, were set up. Mark each item according to the test tube number called for. All substances were dissolved in water. All test tubes were kept at 37 ° C. (water boils at 100° C.). For test tube 5, Substance X was boiled and then cooled before it was added to the protein.

<u>Test Tube</u>	<u>Contents of tubes</u>	<u>Amt. of Substance W present after 24 hours</u>
1	Protein plus Substance X	.05 gram
2	Protein plus Water	.00 gram
3	Protein plus Substance X plus hydrochloric acid	.08 gram
4	Protein plus Hydrochloric acid	.00 gram
5	Protein plus Substance X (boiled)	.00 gram

58. Give the number of the test tube which acts as a control (comparison) for the entire experiment.
59. Give the number of the test tube which gives evidence that protein does not break down spontaneously into Substance W.

60. Give the number of the test tube which gives evidence that Substance X is the active substance in the break down of proteins.
61. Give the number of the tube which is the control for item 60.
62. Give the number of the test tube which shows that a temperature of  $37^{\circ}$  C. does not cause protein to break down into Substance W.
63. Which test tube gives evidence that Substance X is not a stable substance?
64. Which tube is the control for item 63?
65. Which tube gives evidence that acid accelerates the activity of Substance X?
66. Which tube is the control for item 65?
67. Which tube gives evidence that Substance X is a substance whose properties can be destroyed?
68. Give the test tube number of the control for item 67.
69. Which test tube gives evidence that acid affects the protein in some way so that Substance X can act upon it more easily?
70. Give the tube number which is the control for item 69.
71. Give the number of the test tube which indicates that hydrochloric acid alone is ineffective in breaking down proteins.
72. Give the control for item 71.

TABLE XXXXI  
ITEM ANALYSIS DATA FOR TEST F

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*91.1	77.8	.24			
	**88.9	72.2	.24	15	80	68
2	77.8	24.4	.55			
	72.2	5.6	.70	52	38	44
3	64.4	22.2	.44			
	56.6	2.8	.68	50	28	38
4	88.9	60.0	.37			
	86.1	50.0	.41	27	68	60
5	31.1	24.4	.08			
	13.9	5.6	.18	11	10	23
6	100.0	73.3	.62			
	100.0	66.7	.64	46	82	69
7	86.7	51.1	.43			
	83.3	38.9	.48	32	61	56
8	100.0	82.2	.52			
	100.0	77.8	.55	37	88	75
9	88.9	64.4	.34			
	86.1	55.6	.36	23	70	61
10	97.8	80.0	.45			
	97.2	75.0	.47	31	86	73
11	93.3	73.3	.34			
	91.7	66.7	.38	24	79	67
12	95.6	80.0	.35			
	94.4	75.0	.35	24	79	67
13	80.0	51.1	.33			
	75.0	38.9	.36	23	57	54

\* Method of Flanagan

\*\* Method of Davis

TABLE XXXXI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	91.1 88.6	48.8 36.1	.50 .63	38	61	56
15	88.9 86.1	53.3 41.7	.44 .48	32	63	57
16	33.3 16.7	8.9 0.0	.38 .48	32	8	21
17	77.8 72.2	55.6 44.4	.25 .29	18	57	54
18	77.8 72.2	53.3 41.7	.28 .31	19	59	54
19	71.1 63.9	35.5 19.4	.36 .46	30	42	46
20	28.9 11.1	15.6 0.0	.18 .31	19	59	54
21	62.6 52.8	53.3 41.7	.11 .10	6	46	48
22	15.6 0.0	6.7 0.0	.16 .00	0	0	0
23	42.2 27.8	42.2 27.8	.00 .00	0	28	38
24	35.5 19.4	22.2 2.8	.16 .45	29	10	23
25	100.0 100.0	86.7 83.3	.47 .50	33	91	78
26	46.7 33.3	28.9 11.1	.19 .32	20	22	34
27	95.6 94.4	53.3 41.7	.60 .63	45	69	60
28	53.3 41.7	28.9 11.1	.25 .39	25	25	36

TABLE XXXXI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
29	97.8	73.3	.52			
	97.2	66.7	.52	35	82	69
30	93.3	80.0	.25			
	91.7	75.0	.29	18	83	70
31	71.1	33.3	.38			
	63.9	16.7	.51	34	40	45
32	91.1	68.9	.33			
	88.9	61.1	.37	23	74	64
33	77.8	40.0	.38			
	72.2	25.0	.47	31	48	49
34	93.3	71.1	.37			
	91.7	63.9	.39	25	77	66
35	53.3	35.5	.18			
	41.7	19.4	.23	14	31	40
36	64.4	35.5	.29			
	55.6	19.4	.39	25	38	44
37	20.0	6.7	.25			
	0.0	0.0	.00	0	0	0
38	100.0	88.9	.45			
	100.0	86.1	.46	30	93	81
39	71.1	35.5	.36			
	63.9	19.4	.43	25	38	44
40	91.1	35.5	.61			
	88.9	22.2	.67	49	55	53
41	86.7	44.4	.48			
	83.3	30.6	.54	36	57	54
42	62.6	35.5	.27			
	52.8	19.4	.34	21	38	44
43	100.0	62.6	.68			
	100.0	52.8	.71	54	74	64

TABLE XXXXI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
44	93.3	40.0	.62			
	91.7	25.0	.69	51	57	54
45	97.8	80.0	.45			
	97.2	75.0	.46	30	87	72
46	80.0	28.9	.52			
	75.0	11.1	.63	45	42	46
47	91.1	73.3	.29			
	88.9	66.7	.31	19	77	66
48	86.7	86.7	.00			
	83.3	83.3	.00	0	83	70
49	100.0	100.0	.00			
	100.0	100.0	.00	0	100	100
50	0.0	0.0	.00			
	0.0	0.0	.00	0	0	0
51	100.0	86.7	.48			
	100.0	83.3	.50	33	91	79
52	88.9	42.2	.53			
	86.1	27.8	.59	41	57	54
53	62.6	11.1	.55			
	52.8	0.0	.73	56	27	37
54	97.8	55.6	.65			
	97.2	44.4	.68	50	70	61
55	91.1	53.3	.48			
	88.9	41.7	.54	36	64	58
56	100.0	75.6	.58			
	100.0	69.4	.62	44	84	71
57	93.3	64.4	.42			
	91.7	55.6	.47	31	73	63
58	93.3	71.1	.36			
	91.7	63.9	.39	25	77	66

TABLE XXXXI (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
59	97.8	40.0	.72			
	97.2	25.0	.76	60	61	56
60	95.6	46.7	.64			
	94.4	33.3	.67	49	64	58
61	80.0	26.7	.53			
	75.0	8.3	.68	50	40	45
62	95.6	37.7	.67			
	94.4	22.2	.73	56	57	54
63	93.3	35.5	.60			
	91.7	19.4	.72	55	55	53
64	97.8	42.2	.72			
	97.2	27.8	.75	59	61	56
65	100.0	91.1	.40			
	100.0	88.9	.41	27	93	82
66	82.2	44.4	.42			
	77.8	30.6	.48	32	53	52
67	100.0	80.0	.55			
	100.0	75.0	.58	40	86	73
68	100.0	73.3	.62			
	100.0	66.7	.64	46	82	69
69	97.8	88.9	.32			
	97.2	86.1	.32	20	91	79
70	84.4	44.4	.45			
	80.6	30.6	.51	34	55	53
71	100.0	86.7	.45			
	100.0	83.3	.50	33	91	78
72	*91.1	66.7	.35			
	**88.9	58.3	.39	25	73	63

\* Method of Flanagan

\*\* Method of Davis

## TEST G

## DRAWING OF CONCLUSIONS

This test was designed to measure your ability to make conclusions. When facts are analysed and studied they sometimes yield evidence which help in the solution of a problem. However, any conclusion must be checked before it can be accepted. The following key includes four ways in which conclusions may be faulty. Each of the items present a question or problem, a brief description of an experiment and one or more conclusions drawn from the experiment. Each experiment was repeated many times. Read each problem, experiment and the conclusions. Where several conclusions are given evaluate each conclusion separately. Is the conclusion tentatively justified by the data? If so, mark space 1 on your answer sheet. If the conclusion is not justified determine whether 2, 3, 4, or 5 in the key is the best reason for it being faulty and mark the proper space on your answer sheet.

Key

The conclusion is:

1. Tentatively justified.
2. Unjustified - it does not answer problem.
3. Unjustified - the experiment lacks a control (comparison).
4. Unjustified - the data are faulty or inadequate, though a control was included.
5. Unjustified - it is contradicted by the data.

PROBLEM: A student was interested in developing a test for a certain type of substance. In all 100 cases his test was positive.

1. He concluded that the test was a specific test for the substance.

PROBLEM: A student knew that a purple color develops when iodine is added to starch and that this is a specific test for starch. He wished to determine whether a certain food contained starch. He added iodine to the food and found that it turned purple.

2. He concluded that the food was fattening.



3. Another student concluded that iodine is a test for starch.

PROBLEM: An investigator wanted to know what causes people to breathe faster when they are running rapidly. He found that breathing more carbon dioxide increased the breathing rate, but that the breathing of air deficient in oxygen did not increase the breathing rate.

4. He concluded that people breathe faster when they are running because they need more oxygen.
5. Someone else concluded that running increases the rate of breathing.
6. Another person said that people running rapidly take in more carbon dioxide, causing them to breathe more rapidly.
7. Still another claimed that it is harder for the heart to pump faster without sufficient oxygen.
8. Another concluded that carbon dioxide affects the breathing rate.
9. Someone else concluded that people who are exercising must breathe pure carbon dioxide to cause an increase in breathing rate.

PROBLEM: An individual, wishing to determine whether oxygen is used during sleep, analyzed the expired air of a large number of sleeping persons. He found that the expired air contained oxygen.

10. He concluded that oxygen is not used during sleep.
11. Another concluded that oxygen is needed for life.
12. Someone else claimed that people breathe while they are sleeping.
13. Still another person concluded that oxygen is given off as well as taken in during sleep.
14. Another person said that this proved that oxygen is used during sleep.

PROBLEM: An investigator wished to determine whether temperature increased the rate of a certain reaction. On repeated tests he found that if he started out with a certain amount of his original substances he would obtain, after one hour, 1 gram of the substance produced by the reaction at  $0^{\circ}\text{C}.$ , 2 grams at  $20^{\circ}\text{C}.$ , 5 grams at  $40^{\circ}\text{C}.$  and 3 grams at  $60^{\circ}\text{C}.$

15. He concluded that increased temperature increased the rate of the reaction.
16. Another person claimed that this shows that an increase in temperature increases the amount of the original substance.

PROBLEM: A person wanted to determine whether bile aided in the digestion of fats. He found that whenever he mixed pancreatic juice with fats a small part of the fat was digested, but whenever he mixed pancreatic juice and bile with fat, he found that the fat was completely digested. When he mixed bile alone with fat he found that there was no digestion.

17. He concluded that bile aided in the digestion of fats.
18. Another concluded that pancreatic juice was necessary for digestion of fats.
19. One person concluded that it was necessary that the bile and pancreatic juice work together, in order that fats may be digested.
20. Someone else claimed that bile does not aid in the digestion of fat.

PROBLEM: In order to find out if all foods contained starch, ten foods were tested by the iodine test which was known to be a specific test for starch. All of the foods tested contained starch.

21. The conclusion drawn was that all foods contain starch.
22. Another conclusion was that iodine is a good reagent to determine the presence of starch.
23. Another conclusion was that the iodine test proved that starch was present.

PROBLEM: In order to determine whether corticosterone caused a certain disease, a person analyzed the blood of several hundred patients suffering from the disease. He found that in each case the blood contained cortin.

24. He concluded that the disease was caused by corticosterone.

PROBLEM: In order to determine the cause of increased red blood cell count at high altitude, experimenters subjected rats, dogs and guinea pigs at sea level to a reduced total atmospheric pressure. The red cell count was higher in these than in the same kinds of animals not subjected to reduced atmospheric pressure.

25. Conclusion: A decrease in the oxygen in the air breathed at high altitude causes the increase in red cell count.

26. Another conclusion: The red cell count varies inversely with the atmospheric pressure.

PROBLEM: Two students desired to know whether certain types of mosquitos or whether all mosquitos spread malarial fever. They captured many specimens of three kinds of wild mosquitos, types A, B, and C. They examined the digestive tracts of all three types. They found malarial parasites only in type A mosquitos.

27. Conclusion: Malarial fever is spread by type A mosquitos but not by types B and C.

28. Another conclusion: Not all mosquitos carry malaria parasites.

29. Another conclusion: Not all mosquitos have malarial parasites.

PROBLEM: A student interested in frozen food preservation wanted to determine whether extremely low temperatures killed the kind of bacteria that spoil meat. He cut a number of pieces of various types of meat into two pieces leaving one piece of each sample at room temperature and the other of each sample in a locker at a temperature of 40 degrees below freezing. All samples were sealed in bacteria-proof containers. After thirty days he opened the packages. He found the room temperature

## PROBLEM: (continued)

samples badly decomposed. The frozen samples were in their original condition except for being frozen solid.

30. Conclusion: A temperature 40 degrees below freezing will kill the bacteria that are responsible for the decay of meat.
31. Another conclusion: Heat is a controlling factor in the preservation of foods.
32. Another conclusion: Meat kept in a temperature of 40 degrees below freezing does not become decomposed.
33. Another conclusion: Room temperature causes meats to spoil, whereas frozen meats are preserved.
34. Still another conclusion: Bacteria must not have been present in the frozen packages.

PROBLEM: A person wanted to know what caused a certain disease. He examined 1000 patients with the disease. All had a certain bacteria (Bacteria A) in the digestive tract.

35. He concluded that Bacteria A was the cause of the disease.
36. Another conclusion: The disease starts in the digestive tract.
37. Another conclusion: Bacteria A is necessary for digestion.
38. Another conclusion: The cause of the disease was spoilage of food.

PROBLEM: A person wanted to know why plants bend toward the light. He placed one group of plants in the light with the light source at the right. He placed another group of similar plants in the dark. The plants in the dark grew straight, the plants in the light were bent to the right.

39. He concluded that plants bend toward the light.
40. Another concluded that plants bend toward the light because they need light to grow.

41. Someone else concluded that light influences the direction in which plants grow.
42. Another concluded that plants bend toward the sun in order to get the beneficial rays of the sun.

PROBLEM: An investigator wanted to know what caused fish to swim against the current. He placed fish in a bottle. If the bottle was moved to the right the fish moved to the left and vice versa. Blind fish did not respond to the water currents in the bottle, but fish do orient against the current in a stream at night.

43. He concluded that fish can see at night.
44. Another concluded that fish swim against the current because fish will drown if water enters the rear of the gills with force over a long period.
45. Another concluded that normal fish swim against the current.
46. Someone else concluded that blind fish do not swim against the current because they cannot see.

PROBLEM: Investigator A wanted to know what caused people to become ill if confined in large numbers to a small closed area. He found on repeated tests that the air in very crowded closed areas contained about 5% carbon dioxide, while normal air contains .03% carbon dioxide.

47. He concluded that excessive carbon dioxide caused the illness.
48. Another investigator concluded that the illness was caused by insufficient oxygen.
49. Another investigator claimed that the illness was caused by the germs exhaled by the people in the room.

PROBLEM: Investigator B in an attempt to solve the same problem repeated the experiment done by investigator A but in addition had people in uncrowded rooms breathe air containing 5% carbon dioxide. No ill effects were noted among those in the uncrowded rooms.

50. He also concluded that excessive carbon dioxide caused the illness.

51. Another investigator claimed that this showed that the disease was caused by insufficient oxygen.
52. The investigator who calimed the disease was due to germs was convinced by this experiment that he was correct.
53. Another conclusion was that 5% carbon dioxide will produce no ill effects.
54. Still another claimed that people live better in uncrowded areas.

PROBLEM: To find out if all foods contain sugar. Benedict's solution is a specific test for sugar. Ten foods were tested with Benedict's solution. All of the foods contained sugar.

55. Conclusion: Benedict's solution is a good test for sugar.
56. Another conclusion: All foods contain sugar.
57. Another conclusion: The Benedict test showed that sugar was present.

PROBLEM: To determine whether a certain bacteria uses oxygen. The Winkler test is an oxygen test. A broth in which bacteria were grown was tested for oxygen. The broth was shown, by the Winkler test, to contain oxygen.

58. Conclusion: This type of bacteria does not use oxygen.
59. Another conclusion: This type of bacteria gives off oxygen as a waste product.
60. Still another conclusion: The presence of oxygen does not stop the growth of bacteria.
61. Another person concluded that this proves that oxygen is needed by bacteria.

PROBLEM: To determine the cause of disease X. One thousand persons with the disease were examined. Bacteria Q was found in the mouth of all of the persons with the disease.

- 62. One conclusion: Bacteria Q causes the disease.
- 63. Another conclusion: This disease starts in the mouth.
- 64. Another conclusion: This disease is caused by bacteria introduced into the mouth from contaminated food.

PROBLEM: To determine the reaction of insects to light. Flies were placed in a jar, the upper half of which was covered with black paper. A light was placed near the jar. All of the flies flew to the lower half of the jar and toward the illuminated side.

- 65. Conclusion: Insects are attracted to light.
- 66. Another conclusion: Insects are attracted to heat.
- 67. Another conclusion: The flies needed light for warmth.

PROBLEM: To determine some of the requirements for the sprouting of seeds. Two groups of plants were planted in flower pots. Conditions of both were the same except that one pot was put in the greenhouse at 40 degrees; the other group was put in a greenhouse at 70 degrees. Those in the cold room did not sprout, those in the warm room sprouted. Many kinds of seeds were used in each group.

- 68. Conclusion: A temperature of 70 degrees is required for seeds to sprout.
- 69. Another conclusion: Plants need heat to live.
- 70. Another conclusion: Moisture is one of the requirements for the sprouting of seeds.
- 71. Another conclusion: For anything to grow energy is needed.
- 72. Another conclusion: A temperature of 40 degrees keeps seeds from sprouting.

PROBLEM: To determine some of the requirements for the sprouting of seeds. Two groups of seeds were planted. Conditions were the same for both groups except that one group was planted in

## PROBLEM: (continued)

stoppered bottles, the other group in open bottles. Only the seeds in the open bottles sprouted. Many different kinds of seeds were included in each group.

- 73. Conclusion: Seeds require oxygen to sprout.
- 74. Another conclusion: One of the requirements for the sprouting of seeds is moisture.
- 75. Another conclusion: The seeds in the stoppered bottles were dormant.
- 76. Another conclusion: Energy from the outside is necessary for growth.
- 77. Another conclusion: Carbon dioxide is a requirement of sprouting seeds.

PROBLEM: What are some of the requirements for seeds to sprout? A student put many different kinds of seeds in pots containing garden soil and many different kinds of seeds in pots containing the same type of soil with all of the potassium salts removed. The plants in the garden soil grew and developed well. The plants in the other pots were small and soon died. All other conditions were the same for both groups.

- 78. Conclusion: Potassium salts are required for seeds to sprout.
- 79. Another conclusion: Heat and moisture are necessary for seeds to sprout.
- 80. Another conclusion: Minerals are essential for the germination of seeds.
- 81. Another conclusion: Potassium salts contain some important energy for plants.
- 82. Another conclusion: When the plants had used up their supply of food they couldn't replace it.
- 83. Another conclusion: Potassium salts as well as other minerals are essential to plants and their lack will slow down growth.



PROBLEM: What are some of the requirements for seeds to sprout? The student placed two groups of seeds in two pots and watered one pot daily. The other group he watered on alternate days. All of the seeds sprouted. Many types of seeds used, other conditions same for both groups.

84. Conclusion: Water is necessary if seeds are to sprout but it is not necessary to water them every day.
85. Another conclusion: Seeds will sprout with a limited amount of water.
86. Another conclusion: One of the requirements of seeds to sprout is moisture.
87. Another conclusion: Water is a minor factor in the sprouting of seeds.
88. Another conclusion: Both groups of plants had an adequate amount of water.

PROBLEM: What are some of the requirements for seeds to sprout? The same student planted two groups of seeds of different types in pots and placed one group of the pots in the light, the others in the dark. Those plants in the light were green, those in the dark were yellow. Other conditions were the same for both groups.

89. Conclusion: Light is necessary for sprouting of seeds.
90. Another conclusion: Plants require light to mature properly.
91. Another conclusion: Light makes the plants green.

PROBLEM: An investigator wanted to determine whether increased light increased the rate of a certain reaction. On repeated tests it was found that a certain amount of the original substance (X), after one hour, would produce 1 gram of substance Y with 10 photons (units of light) of illumination, 2 grams with 20 photons, 4 grams with 30 photons and 3 grams with 40 photons.

92. Conclusion: Increased amount of light increases the rate of the reaction.

93. Another conclusion: Heat increased the rate of the reaction.

PROBLEM: A student wanted to determine whether plants grow more rapidly in the light or in the dark. Two groups of seeds were planted. After two weeks the plants were measured. Those in the light were green and a few inches long. Those in the dark were yellow and a foot long. All other conditions were the same for both groups. The experiment was repeated with several kinds of seeds. The results were the same as given above.

94. Conclusion: The plants in the dark put all their energy into height trying to reach light while the other ones put their energy into strength.
95. Another conclusion: Light is necessary for faster and better growth of plants.
96. Another conclusion: The plants grown in the light were more healthy.
97. Another conclusion: Plants grow more rapidly in the dark.
98. Another conclusion: Light is necessary for the development of the green color of plants.

PROBLEM: A student wanted to determine whether a certain beverage contained sugar. Benedict's solution which is blue when added to sugar and heated turns the solution yellow. (It is known to be a specific test for sugar). Benedict's was added to the beverage and heated. The solution turned yellow.

99. Conclusion: The beverage is not fattening.
100. Another student concluded that Benedict's solution is a good test for sugar.

TABLE XXXXII

## ITEM ANALYSIS DATA FOR TEST G

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*62.6	31.1	.33			
	**52.8	13.9	.43	28	33	41
2	77.8	64.4	.17			
	72.2	55.6	.15	9	64	58
3	60.0	24.4	.37			
	50.0	5.6	.56	38	28	38
4	95.6	35.5	.72			
	94.4	19.4	.75	59	55	53
5	80.0	40.0	.42			
	75.0	25.0	.50	33	50	50
6	20.0	4.4	.38			
	0.0	0.0	.00	0	0	0
7	37.7	35.5	.02			
	22.2	19.4	.04	2	21	33
8	20.0	8.8	.20			
	0.0	0.0	.00	0	0	0
9	33.3	29.8	.04			
	16.7	11.1	.07	4	14	27
10	46.7	26.7	.21			
	33.3	8.3	.38	24	21	33
11	91.1	60.0	.42			
	88.9	50.0	.46	30	68	60
12	88.9	40.0	.54			
	86.1	25.0	.63	45	57	54
13	22.2	6.7	.32			
	2.8	0.0	.12	7	2	5

\* Method of Flanagan

\*\* Method of Davis

TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	22.2 2.8	13.3 0.0	.14 .12	7	2	5
15	46.7 33.0	17.8 0.0	.33 .65	47	17	30
16	35.5 19.4	26.7 8.3	.13 .23	14	14	27
17	95.6 94.4	68.9 61.1	.47 .50	33	78	66
18	46.7 33.0	4.4 0.0	.58 .65	47	17	30
19	22.2 4.4	2.8 0.0	.37 .12	7	2	5
20	100.0 100.0	68.9 61.1	.65 .67	49	80	68
21	20.0 4.4	24.4 5.6	-.05 -.07	-4	5	15
22	93.3 91.7	33.3 16.7	.65 .72	56	55	53
23	77.8 72.2	8.9 0.0	.69 .81	69	37	43
24	60.0 50.0	37.7 22.2	.23 .31	19	35	42
25	31.1 13.9	13.3 0.0	.25 .46	30	8	20
26	15.6 0.0	6.7 0.0	.18 .00	0	0	0
27	37.7 22.2	28.9 11.1	.10 .18	11	17	30
28	26.7 2.2	8.3 0.0	.50 .35	22	5	14

TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
29	37.7	2.2	.61			
	22.2	0.0	.55	37	12	25
30	26.7	26.7	.00			
	8.3	8.3	.00	0	8	20
31	48.8	11.1	.46			
	36.1	0.0	.66	48	18	31
32	60.0	8.9	.58			
	50.0	0.0	.72	55	25	36
33	62.5	20.0	.44			
	52.8	0.0	.73	56	27	37
34	33.3	6.7	.40			
	16.7	0.0	.48	32	9	21
35	71.1	42.2	.30			
	63.9	27.8	.36	23	46	48
36	60.0	31.1	.30			
	50.0	13.9	.41	27	31	40
37	77.8	53.3	.29			
	72.2	41.7	.31	19	57	54
38	13.3	8.9	.00			
	0.0	0.0	.00	0	0	0
39	91.1	15.6	.72			
	88.9	0.0	.88	82	44	47
40	66.7	24.4	.43			
	58.3	5.6	.61	43	31	40
41	55.6	6.7	.60			
	44.4	0.0	.70	52	22	34
42	42.2	22.2	.23			
	27.8	2.8	.46	30	16	29
43	86.7	33.3	.57			
	83.3	16.7	.67	49	50	50

TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
44	53.3 41.7	35.5 19.4	.18 .23	14	31	40
45	75.6 69.4	11.1 0.0	.65 .68	81	35	42
46	73.3 66.7	17.8 0.0	.56 .79	65	33	41
47	6.7 0.0	0.0 0.0	.35 .00	0	0	0
48	75.6 69.4	17.8 0.0	.58 .81	68	35	42
49	66.7 58.3	42.2 27.8	.26 .31	19	42	46
50	95.6 94.4	66.7 58.3	.49 .46	30	74	64
51	62.6 52.8	17.8 0.0	.48 .73	56	27	37
52	48.8 36.6	24.4 5.6	.26 .43	29	21	33
53	51.1 38.9	8.9 0.0	.51 .67	49	19	32
54	95.6 94.4	26.7 8.3	.74 .83	71	51	51
55	95.6 94.4	26.7 8.3	.74 .83	71	51	51
56	20.0 0.0	20.0 0.0	.00 .00	0	0	0
57	84.4 80.6	0.0 0.0	.87 .84	74	40	45
58	37.7 22.2	13.3 0.0	.32 .55	37	12	25

TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
59	26.7 8.3	6.7 0.0	.34 .35	22	5	14
60	86.7 83.3	35.5 19.4	.55 .63	45	51	51
61	35.5 19.4	22.2 2.8	.15 .39	25	10	23
62	66.7 58.3	40.0 25.0	.28 .34	21	42	46
63	53.3 41.7	37.7 22.2	.16 .23	14	31	40
64	20.0 0.0	22.2 2.8	.00 -.12	7	2	5
65	4.4 0.0	4.4 0.0	.00 .00	0	0	0
66	40.0 25.0	17.8 0.0	.26 .58	40	13	26
67	48.8 36.1	31.1 13.9	.18 .29	18	25	36
68	51.1 38.9	28.9 11.1	.23 .35	22	24	35
69	46.7 33.3	17.8 0.0	.33 .64	46	17	30
70	82.2 77.8	33.3 16.7	.50 .61	43	46	48
71	60.0 50.0	17.8 0.0	.45 .72	55	25	36
72	33.3 16.7	4.4 0.0	.48 .48	32	8	21
73	22.2 2.8	13.3 0.0	.14 .12	7	2	5

TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
74	73.3	35.5	.38			
	35.5	19.4	.46	30	44	47
75	48.8	22.2	.30			
	36.1	2.8	.56	38	19	32
76	44.4	17.8	.31			
	30.6	0.0	.62	44	16	29
77	37.7	17.8	.23			
	22.2	0.0	.55	37	12	25
78	26.7	22.2	.06			
	8.3	2.8	.17	10	5	15
79	80.0	24.4	.56			
	75.0	5.6	.71	54	40	45
80	28.9	8.9	.32			
	11.1	0.0	.40	26	6	16
81	33.3	6.7	.42			
	16.7	0.0	.48	32	8	21
82	73.3	40.0	.34			
	66.7	25.0	.41	27	46	48
83	11.1	2.2	.34			
	0.0	0.0	.00	0	0	0
84	6.7	4.4	.08			
	0.0	0.0	.00	0	0	0
85	20.0	8.9	.20			
	0.0	0.0	.00	0	0	0
86	20.0	4.4	.34			
	0.0	0.0	.00	0	0	0
87	51.1	24.4	.29			
	38.9	5.6	.47	31	22	34
88	55.6	6.7	.58			
	44.4	0.0	.70	52	22	34



TABLE XXXXII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
89	75.6	40.0	.37			
	69.4	25.0	.43	28	48	49
90	35.5	6.7	.45			
	19.4	0.0	.52	35	10	23
91	46.7	4.4	.57			
	33.0	0.0	.64	46	17	30
92	51.1	6.7	.55			
	38.9	0.0	.67	49	19	32
93	53.3	20.0	.36			
	41.7	0.0	.69	51	21	33
94	44.4	4.4	.57			
	30.6	0.0	.62	44	17	30
95	86.7	62.6	.32			
	83.3	52.8	.35	22	68	60
96	77.8	17.8	.60			
	72.2	0.0	.81	69	37	43
97	91.1	91.1	.00			
	88.9	88.9	.00	0	89	76
98	86.7	11.1	.74			
	83.3	0.0	.86	77	40	45
99	77.8	53.3	.28			
	72.2	41.7	.31	19	57	54
100	*97.8	40.0	.72			
	**97.2	25.0	.77	61	61	56

\* Method of Flanagan  
 \*\* Method of Davis

## TEST H

## INTERPRETATION OF DATA

## TEST J

## GENERALIZATIONS AND ASSUMPTIONS

This test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

In freezing of vegetables the common practice for both commercial and home frozen vegetables is to scald the vegetables first, by placing them in boiling water for two or three minutes. The following data were obtained in an experiment which measured the amounts of Vitamin C in fresh vegetables, scalded vegetables before freezing, and vegetables frozen for six months. One group of the frozen vegetables was frozen without first scalding, the other group was first scalded. The Vitamin C content of the frozen vegetables was determined before and after they were cooked. All figures indicate the amount of Vitamin C in mg. per 100 cc.

<u>Vegetable</u>	<u>Fresh</u>	<u>Scalded</u>	<u>Frozen</u>			
			<u>Unscalded</u>		<u>Scalded</u>	
			<u>Raw</u>	<u>Cooked</u>	<u>Raw</u>	<u>Cooked</u>
Chard (greens)	60	37	20	2	24	14
Spinach	82	43	10	1	27	16
Peas	29	21	14	10	20	16
Green beans	34	29	25	13	23	17
Lima beans	33	20	26	18	20	14

1. Scalding of all vegetables causes destruction of some of the Vitamin C content of the vegetables.
2. Spinach is a good source of Vitamin C.
3. Leafy green vegetables are a better source of Vitamin C than the pod type vegetables.
4. Leafy green vegetables are a better source of Vitamin C than root vegetables.
5. The practice of scalding leafy vegetables before freezing should be eliminated because scalding destroys some of the Vitamin C.
6. Lima beans should be frozen without scalding provided the quality of the unscalded product is equal to the scalded in other respects.
7. A better tasting product is obtained if lima beans are scalded before freezing.
8. After commercially frozen peas have been cooked they are a good source of Vitamin C as commercially frozen chard which has been cooked.
9. The percentage of the total Vitamin C destroyed by scalding is about the same for all vegetables.
10. Since the vitamin content of food is an important consideration in its purchase, in buying frozen green vegetables one should be careful in choosing the kind of vegetables because the Vitamin C content of different frozen vegetables varies considerably.
11. The breakdown of Vitamin C is hastened by heating.
12. Since frozen leafy vegetables are much easier to prepare, the practice of using them exclusively is justified from the dietary standpoint.

13. Frozen orange juice contains somewhat less Vitamin C than freshly extracted orange juice.
14. (Fresh spinach is usually cooked for about ten minutes). Cooked spinach (unfrozen) contains less Vitamin C than scalded spinach.
15. Heating causes some change to occur in the Vitamin C molecule.

Items 16 through 21 are a re-evaluation of some of the items 1 through 15. Re-read items 1, 3, 9, 11, 13 and 15 and determine whether they are generalizations, extensions of data, explanations of the data or merely restatements of the data, etc. Answer each according to the following key:

#### Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
  2. The data indicates a trend which if continued in either direction would make the statement true.
  3. An explanation of the data in terms of cause and effect.
  4. A restatement of results.
  5. None of the above.
- 
- |             |              |
|-------------|--------------|
| 16. Item 1. | 19. Item 11. |
| 17. Item 3. | 20. Item 13. |
| 18. Item 9. | 21. Item 15. |

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 1 through 15.

#### Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: The breakdown of Vitamin C proceeds spontaneously but is a relatively slow process at low temperature.

22. Vitamin C is a stable substance.
23. There is order in the universe.
24. Vitamin C is not destroyed by the freezing process.
25. Vitamin C responds in a similar way to the environment no matter what the source of Vitamin C is.
26. The Vitamin C content of all the vegetables studied was reduced after being frozen for six months.
27. All chard is similar in its reactions to the chard studied in this experiment.
28. Vitamin C is gradually destroyed by freezing and is not suddenly destroyed.

Conclusion II: The breakdown of Vitamin C is hastened by heating.

29. All vitamins react in the same way.
30. Vitamin C evaporates when heated.
31. All beans are similar in their reaction to the ones studied in this experiment.
32. Heating causes some change to occur in the Vitamin C molecule.
33. Vitamin C reacts in the same way no matter what the source of the Vitamin C.
34. Pod type vegetables have a basic similarity.

Conclusion III: The Vitamin A content of vegetables is affected by heating.

35. Pod type vegetables have a basic similarity.
36. Vitamin C is gradually destroyed by heating.
37. All vitamins react in a similar way to heat.
38. There is a direct relationship between the amount of Vitamin C and Vitamin A in foods.

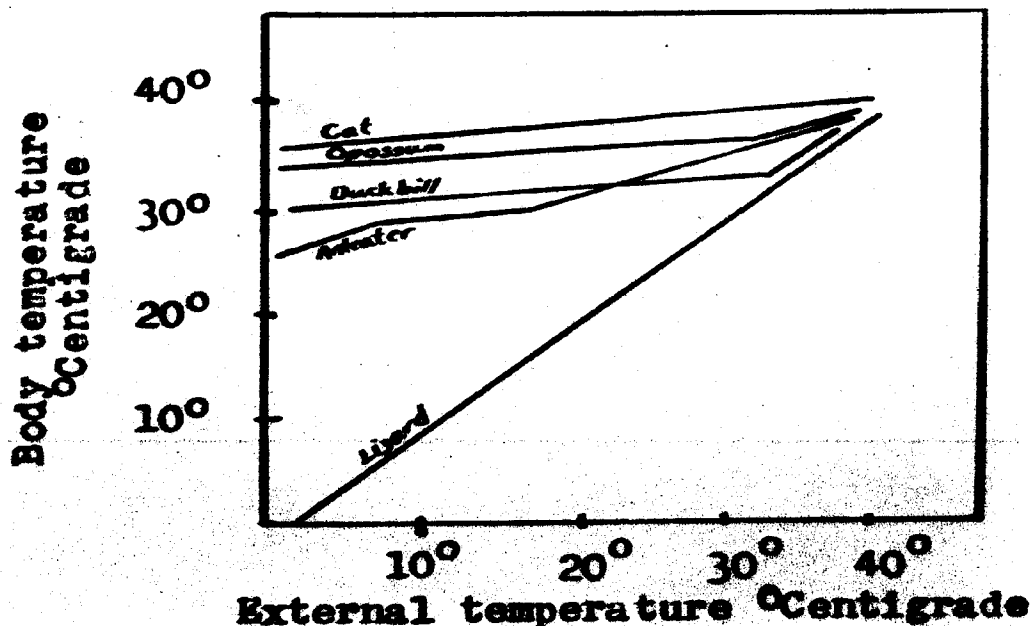
39. There is order in the universe.
40. Heating affected the amount of Vitamin C in the vegetables studied.
41. In all cases studied cooking reduced the Vitamin C content of the vegetables.

This test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

### Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

Items 42 through 61 refer to the following graph. Use the key above to answer the items. The lizard is considered to be cold blooded, the others warm blooded.



42. The body temperature of the cat varies more than the body temperature of the ant eater.
43. The cat and ant eater have some type of mechanism which regulates the body temperature.
44. When the external temperature is  $50^{\circ}\text{C}$ . the temperature of the lizard is also  $50^{\circ}\text{C}$ .
45. The body temperature of warm blooded animals is unaffected by the external temperature.
46. At an external temperature of  $50^{\circ}\text{C}$ . the temperature of the cat is  $50^{\circ}\text{C}$ .
47. When the external temperature is  $50^{\circ}\text{C}$ . the temperature of the ant eater would be higher than the temperature of the cat.
48. The temperature of a mouse would be about half way between that of the cat and the ant eater.
49. At no time during the experiment did any of the animals have the same body temperature.
50. The ant eater exhibits a closer relationship to the lizard than to the opossum.
51. The sharp rise in the body temperature of the lizard indicates that the lizard uses food at a faster rate than the cat.
52. The ability of the cat to maintain its temperature is due to its coat of hair.
53. There is a close correlation between the body temperature of the lizard and that of the external environment.
54. The heart rate of the lizard would increase with temperature in the same way as the body temperature increases.
55. The body temperature of the cat showed the least variation in temperature during the experimental period.
56. The temperature of all of the warm blooded animals was always higher than the external temperature.
57. The warm blooded animals are sufficiently insulated to conserve heat.
58. Warm blooded animals can withstand cold better than cold blooded animals.

59. At 20 degrees below 0°C. the lizard would be frozen.
60. The normal body temperature of the duckbill is higher than that of the echidna.
61. If the temperature of other cold blooded animals were plotted it would resemble that of the lizard.

Items 62 through 68 are a re-evaluation of some of the items 42 through 61. Re-read items 43, 44, 47, 50, 52, 55 and 61 and determine whether they are generalizations, extensions of the data, explanations of the data or merely restatements of the data, etc. Answer each according to the following key:

#### Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
  2. The data indicates a trend which if continued in either direction would make the statement true.
  3. An explanation of the data in terms of cause and effect.
  4. A restatement of results.
  5. None of the above.
- 
- |              |              |
|--------------|--------------|
| 62. Item 43. | 66. Item 52. |
| 63. Item 44. | 67. Item 55. |
| 64. Item 47. | 68. Item 61. |
| 65. Item 50. |              |

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 41 through 61.

#### Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.



Conclusion I: Warmblooded animals have some type of heat regulating mechanism.

- 69. All cats react similarly to changes in temperature.
- 70. It is possible for animals to have some type of heat regulating mechanism.
- 71. The cat and the duckbill are very different in their reaction to the external environment.
- 72. A man and a cat react similarly to the external temperature.
- 73. The lizard has no heat regulating mechanism.
- 74. The opossum had a lower body temperature than the cat.

Conclusion II: Anteaters and duckbills are more closely related than anteaters and cats.

- 75. Similarity of reaction of living things indicate a relationship.
- 76. All anteaters react similarly to changes in external temperature.
- 77. The temperature of the anteater varied more with the external temperature than did that of the cat.
- 78. The degree of closeness of similarity of response of living things runs parallel with the closeness of kinship.
- 79. Close relationship means that two living things have a common ancestor.
- 80. The temperature of the cat varied less than that of the anteater and duckbill with change of temperature.

This test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

Analyses were made of the Vitamin C content of red ripe and green tomatoes as soon as they were picked. Mature green tomatoes were stored at the temperatures indicated in the following table. Those which had ripened by the end of the first week were analyzed for their Vitamin C content; those ripened at the end of the second week were analyzed at the end of the second week, etc. In addition some mature green tomatoes were analyzed each week.

<u>Condition when taken from field</u>	<u>Temp. when stored</u>	<u>No. of weeks stored</u>	<u>Stage of ripeness when analyzed</u>	<u>Vitamin C mg/100 grams</u>
mature green	not stored	0	mature green	15.0
red ripe	not stored	0	red ripe	16.2
mature green	70°F.	1	red ripe	14.4
mature green	70°F.	2	red ripe	12.9
mature green	70°F.	3	red ripe	8.2
mature green	80°F.	1	red ripe	14.0
mature green	80°F.	2	red ripe	9.8
mature green	80°F.	3	red ripe	7.1
mature green	70°F.	1	mature green	10.0
mature green	70°F.	2	mature green	7.2

81. At the time of harvest the green tomatoes were only slightly lower in Vitamin C content than the red ripe ones.
82. Tomatoes which ripened during the first week of storage were almost as high in Vitamin C as those which were ripe at the time of harvest..
83. Tomatoes ripening during the second week of storage were lower in Vitamin C content than those which ripened during the first week.

84. Tomatoes ripened at 90°C. would have less Vitamin C after three weeks than those stored at 80°F.
85. Tomatoes could not be stored at 90°F. because at this high a temperature they would rot or spoil.
86. The lower the temperature at which tomatoes are stored the less is the breakdown of Vitamin C.
87. At 75°F. there would be about 14 mg/100 grams of Vitamin C after a week of storage.
88. Heat causes a breakdown of the Vitamin C molecule.
89. If tomatoes are to be stored for a considerable length of time they should be held at as low a temperature as possible, but high enough to avoid freezing.
90. When one buys tomatoes in the winter the Vitamin C content of the tomatoes compares favorably with the Vitamin C content of those bought fresh in the summer.
91. After four weeks of storage tomatoes stored at 70°F. would contain less than 7 mg/100 grams of Vitamin C.
92. Vitamin C does not develop in the tomatoes as they change from mature green to red ripe on the vine.
93. Some mature green tomatoes ripen in storage within a week.
94. (Tomatoes are often picked green and allowed to ripen during the early fall). The Vitamin C content of these tomatoes is about the same as when they were picked.
95. The green tomatoes which did not ripen in a week had lost about the same amount of Vitamin C as those which ripened during the week..
96. Vitamin C breaks down spontaneously at room temperature.
97. The Vitamin C content of other vegetables decreases if stored at high temperatures.
98. Boiling of vegetables destroys some of the Vitamin C.
99. Vitamin C is a stable substance.
100. Vitamin C is manufactured some place else in the plant than in the fruit (tomato) and is stored in the fruit.

Items 101 through 107 are a re-evaluation of some of the items 81-100. Re-read items 82, 84, 86, 88, 91, 93, and 97 and determine whether they are generalizations, extensions of the data, explanations of the data or merely restatements of the data, etc. Each of these items is to be answered according to the following key:

Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of cause and effect.
4. A restatement of results.
5. None of the above.

- |               |               |
|---------------|---------------|
| 101. Item 82. | 105. Item 91. |
| 102. Item 84. | 106. Item 93. |
| 103. Item 86. | 107. Item 97. |
| 104. Item 88. |               |

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 81 through 100.

Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: Sunlight causes an increase in the Vitamin C content of tomatoes as they ripen on the vine.

108. The test used to measure the amount of Vitamin C in this experiment was a specific test for Vitamin C.

109. The increase of Vitamin C in tomatoes ripening on the vine was caused by the action of sunlight on the leaves.
  110. The tomatoes which were analyzed when green ripe would have contained more Vitamin C if they had been allowed to ripen on the vine.
  111. The test used to measure the amount of Vitamin C accurately measures the amount.
  112. The same results would not have been obtained if the plants had been kept in the dark for the week during which the tomatoes ripened.
  113. All tomatoes would yield the same type of results as those obtained in this experiment.
  114. The Vitamin C content of the tomatoes used in this experiment increased as the tomatoes ripened on the vines.
  115. The Vitamin C was formed in the roots and was transported to the fruits.
  116. The Vitamin C content of ripe tomatoes on the vine was higher than the Vitamin C content of the green ripe tomatoes on the vine.
  117. The plant is capable of manufacturing Vitamin C.
  118. Some change takes place in the Vitamin C molecule at high temperatures.
- Conclusion II: Vitamin C breaks down spontaneously at room temperature.
119. Vitamin C reacts similarly in all plants in which it is found.
  120. Tomatoes are all similar in the amount of Vitamin C they contain.
  121. The Vitamin C content of all tomatoes would decrease when stored at room temperature.
  122. When the tomatoes were stored at room temperature the Vitamin C content decreased.
  123. All vitamins react similarly to storage at room temperature.
  124. There is order in the universe.

125. Vitamin C evaporates at room temperature.
126. The Vitamin C molecule undergoes changes which change the properties of the substance.

This test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

### Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

The following data is concerned with the temperature at which various seeds germinate (sprout). Three kinds of seeds were used, seeds from Species A, Species B and Species C. The number of seeds germinating at various temperature in two weeks is given in the table. No seeds germinated at temperatures below 40°F. or above 95°F.

### Temperatures in Degrees Fahrenheit

	35°	40°	45°	50°	55°	60°	65°	70°	75°	80°	85°	90°	95°	100°
A	0	0	0	0	0	0	5	18	50	70	84	65	30	0
B	0	6	20	41	70	92	65	30	5	0	0	0	0	0
C	0	0	0	4	16	43	72	90	81	52	34	6	0	0

127. Plant B should be planted early in the spring but not in midsummer in middle western states, such as Illinois, Iowa, etc.

128. Plant C is a tropical plant.
129. More seeds of Plant A will germinate at 82° than at any other temperature.
130. None of the seeds of Plant A will germinate below 65°.
131. Seeds do not germinate at freezing temperature.
132. The higher the temperature, the more seeds will germinate.
133. One would not get a crop from plants of the A type in the climate of the northern states, such as Michigan, Minnesota, etc.
134. The optimum temperature for the growth of plants of the C type is 70°.
135. Some seeds of the C variety will germinate at 95°.
136. The optimum temperature for the germination of seeds of the B type is about 56°.
137. Plants of the A type are found in hot wet climates..
138. The rate at which seeds germinate is affected by the temperature.
139. A decrease in moisture reduces the number of seeds germinating more than does a decrease in temperature.
140. If Plant B takes a relatively long time to mature, seeds should be started in greenhouses and set out later if a crop of this type plant is desired in northern states.
141. Plant A could be watermelon.
142. No plants germinate at temperatures above 100°.
143. More seeds would have germinated at lower temperatures if they had been left for a longer time.
144. An increase of 10° above 85° resulted in a much greater reduction in the number of type A seeds germinating than did a reduction of 10°.
145. If one were desirous of raising all three of these plants in one greenhouse one should keep the greenhouse at about 72°.

146. A temperature of  $100^{\circ}$  will kill plants of the B and C types.

Items 147 through 151 are a re-evaluation of some of the items 127 through 146. Re-read items 131, 138, 139, 142 and 144 and determine whether they are generalizations, extensions of the data, interpretations of the data or merely restatements of the data, etc. Each of these items is to be answered according to the following key:

### Key

- Key
1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
  2. The data indicates a trend which if continued in either direction would make the statement true.
  3. An explanation of the data in terms of cause and effect.
  4. A restatement of results.
  5. None of the above.
- 
- |                |                |
|----------------|----------------|
| 147. Item 131. | 150. Item 142. |
| 148. Item 138. | 151. Item 144. |
| 149. Item 139. |                |

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 127-146.

## Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: Seeds will germinate only in the range of temperature from 35°F. to 100°F.

152. The seeds used in this experiment are representative of the extremes of germinating temperatures of seeds.



- 153. No seeds of Species B ever germinate below 35°F.
- 154. None of the seeds which were planted of Species A germinated above 100°F.
- 155. Too few seeds were used in the experiment to make it valid.
- 156. All seeds of Species A behave similarly in their response to temperature to the ones used in this experiment.
- 157. The seeds from Species C germinated at a higher temperature than the seeds of Species B.
- 158. Plants which do not germinate at high temperatures will not grow at high temperatures even when germinated at lower temperatures.
- 159. Seeds will germinate only in a limited temperature range.

Conclusion II: Some seeds of Species B will germinate at 80°F.

- 160. The seeds used in this experiment are completely representative of seeds of Species B.
- 161. A larger sample would yield a greater range of germination temperature.
- 162. All seeds of a species are exactly alike in their response to temperature.
- 163. Some seeds of C germinate at 80°F.
- 164. The entire range in which seeds of Species B will germinate is not represented by this experiment.
- 165. Species B is a cold climate plant.

TABLE XXXXIII  
ITEM ANALYSIS DATA FOR TEST H

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*40.0	20.0	.24			
	**25.0	0.0	.58	40	13	26
2	37.7	24.4	.16			
	22.2	5.6	.31	19	14	27
3	60.0	44.4	.17			
	50.0	30.6	.20	12	40	45
4	91.1	48.8	.51			
	88.9	36.1	.56	38	61	56
5	48.8	13.3	.42			
	36.1	0.0	.66	48	18	31
6	64.4	33.3	.32			
	55.6	16.7	.42	27	35	42
7	95.6	71.1	.45			
	94.4	63.9	.45	29	79	67
8	62.6	37.7	.26			
	52.8	22.2	.32	20	37	43
9	71.1	51.1	.21			
	63.9	38.9	.26	16	51	51
10	17.7	2.0	.43			
	0.0	0.0	.00	0	0	0
11	44.4	20.0	.28			
	30.6	0.0	.62	44	16	29
12	6.7	2.2	.22			
	0.0	0.0	.00	0	0	0
13	22.2	8.9	.23			
	8.9	0.0	.18	11	3	10

\* Method of Flanagan  
\*\* Method of Davis

TABLE XXXXIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	11.1 0.0	15.6 0.0	-.07 .00	0	0	0
15	64.4 55.6	27.2 8.6	.38 .55	37	31	40
42	91.1 88.9	71.1 63.9	.33 .34	21	76	65
43	51.1 38.9	37.7 22.2	.15 .20	12	30	39
44	71.1 63.9	24.4 5.6	.46 .65	47	35	42
45	60.0 50.0	31.1 13.9	.30 .41	27	31	40
46	82.2 77.8	20.0 0.0	.62 .83	72	38	44
47	66.7 58.3	35.5 19.4	.32 .39	25	38	44
48	100.0 100.0	84.4 80.6	.50 .54	36	89	76
49	88.9 86.1	46.7 33.3	.48 .55	37	59	55
50	26.7 8.3	8.9 0.0	.28 .35	22	5	14
51	71.1 63.9	68.9 61.1	.02 .01	1	61	56
52	13.3 0.0	22.2 2.8	-.18 -.12	7	2	5
53	91.1 88.9	64.4 55.6	.38 .40	26	71	62
54	82.2 77.8	77.8 72.2	.04 .10	6	76	65
55	95.6 94.4	77.8 72.2	.37 .38	24	83	70

TABLE XXXXIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
56	57.8 47.2	31.1 13.9	.28 .38	24	30	39
57	44.4 30.6	26.7 8.3	.20 .35	22	19	32
58	33.3 22.2	16.7 2.8	.14 .35	22	9	22
59	66.7 58.3	28.9 11.1	.38 .52	35	35	42
60	0.0 0.0	0.0 0.0	.00 .00	0	0	0
61	66.7 58.3	46.7 33.3	.20 .26	16	44	47
81	88.9 86.1	80.0 75.0	.15 .17	10	80	68
82	75.6 69.4	66.7 58.3	.12 .12	7	63	57
83	93.3 91.7	80.0 75.0	.25 .29	18	83	70
84	80.0 75.0	31.1 13.9	.50 .62	44	42	46
85	95.6 94.4	68.9 61.1	.35 .47	31	77	66
86	57.8 47.2	24.4 5.6	.35 .54	36	27	37
87	62.6 52.8	40.0 25.0	.23 .31	19	38	44
88	71.1 63.9	24.4 5.6	.47 .65	47	35	42
89	60.0 50.0	37.7 22.2	.23 .31	19	35	42
90	42.2 27.8	46.7 33.0	-.05 -.04	-2	44	47

TABLE XXXXIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
91	86.7	53.3	.40			
	83.3	41.7	.43	28	61	56
92	17.8	8.9	.17			
	0.0	0.0	.00	0	0	0
93	97.8	73.3	.48			
	97.2	66.7	.52	35	82	69
94	40.0	11.1	.38			
	25.0	0.0	.58	40	13	26
95	64.4	33.3	.32			
	55.6	16.7	.41	27	35	42
96	40.0	20.0	.24			
	25.0	0.0	.58	40	13	26
97	28.9	11.1	.27			
	11.1	0.0	.40	26	5	17
98	37.7	22.2	.19			
	22.2	2.8	.47	31	12	25
99	46.7	13.3	.40			
	33.0	0.0	.64	46	17	30
100	73.3	42.2	.32			
	66.7	27.8	.38	24	47	48
127	64.4	15.6	.50			
	55.6	0.0	.75	58	28	38
128	64.4	13.3	.54			
	55.6	0.0	.75	58	28	38
129	28.9	11.1	.27			
	11.1	0.0	.40	26	5	17
130	6.7	4.4	.10			
	0.0	0.0	.00	0	0	0
131	37.7	15.6	.30			
	22.2	0.0	.55	37	12	25

TABLE XXXXIII (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
133	51.1 38.9	28.9 11.1	.23 .36	23	25	36
134	4.4 0.0	6.7 0.0	-.10 .00	0	0	0
135	2.2 0.0	6.7 0.0	-.20 .00	0	0	0
136	37.7 22.2	8.9 0.0	.40 .55	37	12	25
137	88.9 86.1	80.0 75.0	.15 .17	10	80	68
138	8.9 0.0	11.1 0.0	.05 .00	0	0	0
139	100.0 100.0	75.6 69.4	.55 .62	44	83	70
140	53.3 41.7	20.0 0.0	.36 .69	51	21	33
141	20.0 0.0	24.4 5.6	-.06 -.27	-17	4	12
142	33.3 16.7	13.3 0.0	.28 .50	33	8	21
143	13.3 0.0	13.3 0.0	.00 .00	0	0	0
144	8.9 0.0	11.1 0.0	.05 .00	0	0	0
145	66.7 58.3	15.6 0.0	.54 .76	60	30	39
146	*51.1 **38.9	6.7 0.0	.46 .67	49	19	32

\* Method of Flanagan  
 \*\* Method of Davis

TABLE XXXIV  
ITEM ANALYSIS DATA FOR TEST J

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
16	*84.4	60.0	.30			
	**80.6	50.0	.34	21	64	58
17	44.4	20.0	.27			
	30.6	0.0	.62	44	16	29
18	48.8	28.9	.22			
	36.1	11.1	.34	21	22	34
19	33.3	33.3	.00			
	16.7	16.7	.00	0	17	30
20	66.7	64.4	.04			
	58.3	55.6	.02	1	57	54
21	46.7	24.4	.25			
	33.3	5.6	.41	27	19	32
22	82.2	64.4	.23			
	77.8	55.6	.24	15	66	59
23	11.1	6.7	.10			
	0.0	0.0	.00	0	0	0
24	35.5	17.8	.23			
	16.7	0.0	.50	33	8	21
25	48.8	17.8	.35			
	36.1	0.0	.66	48	18	31
26	53.3	24.4	.32			
	41.7	5.6	.50	33	24	35
27	48.8	22.2	.30			
	36.1	2.8	.57	39	19	32
28	28.9	4.4	.45			
	11.1	0.0	.40	26	6	17

\* Method of Flanagan  
\*\* Method of Davis

TABLE XXXIV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
29	80.0	60.0	.24			
	75.0	50.0	.26	16	61	56
30	13.3	17.8	-.10			
	0.0	0.0	.00	0	0	0
31	51.1	15.6	.39			
	38.9	0.0	.67	48	19	32
32	20.0	11.1	.16			
	0.0	0.0	.00	0	0	0
33	53.3	24.4	.32			
	41.7	5.6	.50	33	24	35
34	15.6	20.0	-.04			
	0.0	0.0	.00	0	0	0
35	26.7	8.9	.29			
	8.3	0.0	.35	22	5	14
36	64.4	57.8	.07			
	55.6	47.2	.07	4	51	51
37	51.1	24.4	.29			
	38.9	5.6	.48	32	22	34
38	55.6	57.8	-.02			
	44.4	47.2	-.02	- 1	45	47
39	15.6	8.9	.09			
	0.0	0.0	.00	0	0	0
40	64.4	24.4	.42			
	55.6	5.6	.59	39	35	42
41	46.7	8.9	.47			
	33.3	0.0	.64	46	17	30
62	53.3	28.9	.25			
	41.7	11.1	.39	25	28	36
63	62.9	35.5	.28			
	52.8	19.4	.36	23	35	42
64	68.9	37.7	.33			
	61.1	22.2	.40	26	40	45



TABLE XXXIV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
65	40.0	20.0	.24			
	25.0	0.0	.58	40	13	26
66	71.1	42.2	.30			
	63.9	27.8	.36	23	46	48
67	86.7	37.7	.54			
	83.3	22.2	.61	43	53	52
68	84.4	44.4	.44			
	80.6	30.6	.51	34	55	53
69	73.3	51.1	.23			
	66.7	38.9	.29	18	51	51
70	77.8	53.3	.27			
	72.2	41.7	.31	19	57	54
71	53.3	24.4	.32			
	41.7	5.6	.50	33	24	35
72	66.7	46.7	.23			
	58.3	33.3	.26	16	44	47
73	20.0	6.7	.27			
	0.0	0.0	.00	0	0	0
74	71.1	20.0	.52			
	63.9	0.0	.78	63	31	40
75	84.4	46.7	.43			
	80.6	33.3	.48	32	57	54
76	48.8	31.1	.18			
	36.1	13.9	.30	18	25	36
77	64.4	22.2	.43			
	55.6	2.8	.68	50	28	38
78	57.8	20.0	.40			
	47.2	0.0	.70	53	24	35
79	15.6	11.1	.10			
	0.0	0.0	.00	0	0	0
80	66.7	26.7	.41			
	58.3	8.3	.58	40	33	41

TABLE XXXIV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
101	84.4	24.4	.60			
	80.6	5.6	.75	59	44	47
102	82.2	35.5	.48			
	77.8	19.4	.59	41	48	49
103	26.7	24.4	.03			
	8.3	5.6	.07	4	7	19
104	60.0	15.6	.48			
	50.0	0.0	.73	56	25	36
105	77.8	42.2	.38			
	72.2	27.8	.43	28	50	50
106	86.7	35.5	.56			
	83.3	19.4	.63	45	51	51
107	57.8	47.2	.33			
	26.7	8.3	.51	34	28	38
108	64.4	40.0	.25			
	55.6	25.0	.32	20	40	45
109	13.3	11.1	.04			
	0.0	0.0	.00	0	0	0
110	73.3	35.5	.38			
	66.7	19.4	.47	31	42	46
111	62.6	28.9	.34			
	52.8	11.1	.48	32	31	40
112	66.7	40.0	.27			
	58.3	25.0	.34	21	42	46
113	64.4	35.5	.30			
	55.6	19.4	.39	25	38	44
114	37.7	35.5	.03			
	22.2	19.4	.05	3	21	33
115	53.3	22.2	.33			
	41.7	2.8	.59	41	22	34
116	55.6	17.8	.45			
	44.4	0.0	.70	52	22	34

TABLE XXXIV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
117	60.0	35.5	.25			
	50.0	19.4	.35	22	35	42
118	62.6	33.3	.29			
	52.8	16.7	.40	26	33	41
119	53.3	26.7	.28			
	41.7	8.3	.45	29	25	36
120	64.4	44.4	.22			
	55.5	30.6	.27	17	42	46
121	71.1	48.8	.24			
	63.9	36.1	.29	18	50	50
122	60.0	17.8	.45			
	50.0	0.0	.72	55	25	36
123	82.2	46.7	.40			
	77.8	33.3	.46	30	55	53
124	20.0	6.7	.38			
	0.0	0.0	.00	0	0	0
125	17.8	15.6	.04			
	0.0	0.0	.00	0	0	0
126	13.3	6.7	.06			
	0.0	0.0	.00	0	0	0
147	53.3	37.7	.16			
	41.7	22.2	.23	14	31	40
148	40.0	28.9	.12			
	25.0	11.1	.22	13	18	31
149	82.2	42.2	.43			
	77.8	27.8	.50	33	53	52
150	55.6	31.1	.26			
	44.4	13.9	.36	23	28	38
151	8.9	8.9	.00			
	0.0	0.0	.00	0	0	0

TABLE XXXXIV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
152	75.6	53.3	.25	18	55	53
	69.4	41.7	.29			
153	57.8	53.3	.05	3	44	47
	47.2	41.7	.05			
154	53.3	37.7	.16	14	31	40
	41.7	22.2	.23			
155	68.9	28.9	.41	38	35	42
	61.1	11.1	.56			
156	84.4	35.5	.52	42	50	50
	80.6	19.4	.60			
157	64.4	17.8	.48	58	28	38
	55.6	0.0	.75			
158	80.0	51.1	.33	23	57	54
	75.0	38.9	.36			
159	42.2	20.0	.26	43	15	28
	27.8	0.0	.61			
160	68.9	24.4	.44	46	33	41
	61.1	5.6	.64			
161	77.8	33.3	.46	38	44	47
	72.2	16.7	.56			
162	64.4	28.9	.36	43	33	41
	55.5	11.1	.61			
163	22.2	13.3	.14	9	3	10
	2.8	0.0	.15			
164	75.6	15.6	.60	68	35	42
	69.4	0.0	.81			
165	*17.8	11.1	.13	0	0	0
	** 0.0	0.0	.00			

\* Method of Flanagan

\*\* Method of Davis

## APPENDIX II

## TEST I

## THE ABILITY TO THINK SCIENTIFICALLY

## GENERAL DIRECTIONS

1. Place your name, age and sex in the spaces provided on the answer sheet.
2. Place your student number in the space provided for "data of birth".
3. On the space marked "school" place your major.
4. In the space marked "1" below "school" give courses you have had in science in high school, in the space marked "2" give any courses you have had in science in college in addition to biological science.
5. Answer all items; if you don't know - guess.
6. Do not mark on the test booklet. Use scratch paper if you wish.
7. Be sure to mark dark on the answer sheet; the machine does not pick up light markings.
8. Each item has only one answer; do not mark more than one.

This test has been devised to measure your ability to think scientifically. It is divided into several parts, each of these parts tests a different phase of scientific thinking.

This portion of the test is designed to measure your ability to differentiate phases of thinking. These steps include major problems or perplexities, possible solutions to problems, observations which are not results of experimentation but rather preliminary observations, results of experimentation, and conclusions.

The following key is to be used for the succeeding paragraph. Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space in the answer sheet.

#### Key

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Result of experimentation.
4. Initial observation (not experimental).
5. Conclusion (probable solution of problem).

(1) How does a homing pigeon navigate over territory it has never seen before? (2) Do air currents stimulate the pigeon in some way? (3) Are the pigeons equipped with some sort of magnetic compasses; that is, are they sensitive to the earth's magnetism? Yeagley tested the latter by fastening small magnets to the wings of well-trained pigeons. (4) Most of these birds never got home. (5) Others, carrying equal wing weights of non-magnetic copper, made the home roost without trouble, (6) indicating that the earth's magnetism is a factor in pigeon navigation. But the pigeons magnetic compass could not, by itself, bring him back to his roost; because many places on the earth's surface have identical magnetic conditions. Leagley endeavored (7) to determine the other guiding factor. (8) It might be the

sun or stars, but pigeons navigate

# Abbreviated Key

under clouds. While looking at a

map which had lines representing

the intensity of the earth's mag-

1. A major problem

2. Hypothesis

3. Results

4. Observations

5. Conclusions

netism, he noted that the lines were crossed at varying

angles by the parallels of latitude. (9) If pigeons are

sensitive to some factor connected with the lines of lati-

tude, they would have all they need to find their way home.

The next step was (10) to find some physical force, some-

thing the pigeons might be able to detect, related to the

lines of latitude. The effect of the earth's turning varies

directly with latitude; objects near the equator are carried

daily around the earth's circumference, moving at over 1,000

mi. per hr. Objects near the poles are carried around more

slowly. The direction and variation of this circling can be

recorded by various man-made instruments. (11) Why should

not the pigeons feel it, too? (12) If they could, they

would have, along with their magnetic compass, a satisfactory

navigating instrument. Yeagley trained hundreds of pigeons

to return to their home roosts at State College, Pa. Then

he took them to a part of Nebraska where the lines represent-

ing the earth's magnetism cross the parallels of latitude at

the same angle as at State College, Pa. He released the

pigeons to the east of this spot. (13) The pigeons all flew

west. Yeagley believes that (14) pigeons are guided by both

the earth's magnitude and by its turning. (15) Just where

the birds keep their instruments is still unknown; but he

found that (16) birds have a mysterious organ in their eyes,



at the end of the optic nerve. (17) This organ may contain the nerve fibers that pick up vibrations of magnetism and the even more delicate sense that measure the earth's turning.

This portion of the test is designed to test your ability to delimit a problem. A problem is presented. This is followed by a series of questions. Rate the questions according to the following key.

#### Key

1. This question must be answered in order to solve the problem.
2. This question if answered might be useful in the solution of the problem.
3. The answer to this question, though related to the problem, would not help in the solution of the problem.
4. This question is completely unrelated to the problem.
5. This question if answered in the affirmative is a basic assumption of the problem.

PROBLEM: What causes colds?

#### QUESTIONS:

18. Do all people have colds?
19. Is it possible to determine the cause of a cold?
20. Does aspirin help to cure a cold?
21. Can some germ be isolated which, when injected, will cause a cold?
22. Do colds have a cause?
23. Does getting one's feet wet cause a cold?
24. Does becoming chilled after being overheated cause a cold?
25. Why are colds more prevalent in the winter than in the summer?
26. Do other animals get colds?
27. Are people who are tired more susceptible to colds?

**PROBLEM:** What is the function of the thymus gland?  
 (The thymus gland is located in the chest cavity just above the heart.) This gland is largest during the growing period and becomes progressively smaller after maturity.

**QUESTIONS:**

28. Is the gland inactive after maturity?
29. Does the gland have a function?
30. Can any substance be extracted from the gland which when injected into another animal cause growth?
31. If the gland is removed will the animal mature?
32. Can the function of the gland be determined?
33. What causes the gland to grow smaller?

This portion of the test is designed to measure your ability to recognize faulty experimental procedures. In each case a problem and a possible solution to the problem (an hypothesis) are presented. In each case the experiments were designed by students to test the hypotheses. Judge each experiment according to the following key.

Key

This experiment is:

1. Satisfactory
2. Unsatisfactory because it lacks a control or comparison.
3. Unsatisfactory because the control or comparison is faulty.
4. Unsatisfactory because it is unrelated to the hypothesis.
5. None of the above - the experiment is unsatisfactory for reasons other than listed in 2, 3, and 4.

**PROBLEM:** What are some of the requirements for the sprouting of seeds?

**HYPOTHESIS:** Oxygen is a requirement for the sprouting of seeds.

34. If a seed lacked oxygen under a controlled experiment the seed would not function properly and would soon die.

35. Take two packages of seeds. Allow oxygen to be in contact with one package but keep the other package of seeds protected from all oxygen. Observe which sprouts.

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to hypothesis
5. None of the above

36. Place growing plants in an air tight container. Pump out the oxygen. Place other growing plants in containers with oxygen. Keep temperature, light, etc., the same for each.
37. Plant seeds in a container with glass covering it so that no oxygen can enter and see if they sprout. Keep temperature, light and moisture normal.

PROBLEM: A minute insect (aphid) is suspected of spreading a virus disease of roses. How would you determine whether this is true?

HYPOTHESIS: The aphid spreads a virus disease of roses.

38. Put the insect among other kinds of plants other than roses. Leave another group of these plants free from contact with the aphids. Compare the results.
39. Since aphids travel through the air, a plot of roses must be entirely protected from them, and another exposed to aphids which in turn have been exposed to roses afflicted with the virus disease. All must be under constant conditions of soil, atmosphere, etc.
40. Take sample rose with the virus disease. Obtain same kind of rose with no disease. Use microscope to aid in detection of the disease. Use some sort of spray. Note results.
41. Use rose plants which are known not to be diseased. In the same area place rose plants which are diseased but which have been treated to destroy the aphid. Note whether the disease still spreads after the aphids have been killed.
42. In order to determine whether the aphid spreads a virus disease in roses, a group of roses should be put in a hot house free from aphids to see whether they get such a virus disease.

**PROBLEM:** To determine the cause of illness which appears when large numbers of people being confined to a small space.

Abbreviated Key

1. Satisfactory
2. Lacks control
3. Control faulty
4. Unrelated to

**HYPOTHESIS:** Lack of oxygen causes the people to become ill. hypothesis

5. None of the above

43. One might check the oxygen by placing a number of people in a confined place where there was a control amount. Other checks would have to be made also such as the purity of food, the purity of water and whether or not proper sanitation rules were followed.
44. Confine one group to a small space in which there is a limited supply of oxygen. Let the other group have unlimited supply of oxygen and a large space. Let their diets and other items be the same. If the cause of the illness is as stated the confined group will be ill from lack of oxygen.
45. Set two groups of people, one with plenty of oxygen and the other in a normal environment. Determine which group becomes ill.
46. Put a lot of rabbits in a small space for a period of time. Put a few rabbits in the same amount of space. Observe the rabbits and draw conclusions.
47. Put one group of people in a room with an excessive amount of carbon dioxide and another group in a room with a normal amount of carbon dioxide. Keep the oxygen concentration the same in both rooms.

This portion of the test is designed to test your ability to organize data. Select from the key below the curve which best fits the data. If none of the curves fit the data mark space five on your answer sheet. The curves need not have the same amount of slope as the curves presented in the key. Use scratch paper if you wish.

Key

1.



3.



2.






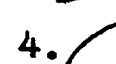
4.



5. None of the curves

48. The horizontal axis represents the time in hours after the injection of sugar into the blood; the vertical axis is the amount of sugar in the blood.

Abbreviated Key

1.  3.   
2.  4.  5. none

Time after injection

1  
3  
6

Blood sugar

35  
12  
8

49. The horizontal axis represents age in years. The vertical axis is the percent increase in the weight of the ovaries and other female sex organs from birth to 20 years.

Age

4  
10  
14  
18

Percent increase

8  
12  
20  
80

50. The horizontal axis represents time in days; the vertical axis is the number of yeast cells in millions (starting with 100 yeast cells).

<u>Time in days</u>	<u>Number of yeast cells in millions</u>
4	25
8	150
12	390
20	400

51. The horizontal axis represents the amount of thyroprotein fed daily to cows. The vertical axis represents the percent increase in milk production.

Thyroprotein fed

.15 grams  
.20 grams  
.24 grams  
.30 grams

Percent increase

18  
23  
27  
33

This test is designed to measure your understanding of the relation of facts to the solution of a problem. The over-all problem involved in this test is presented. This is followed by a series of possible solutions to the problem (hypotheses). After each hypothesis there are a number of items, all of which are true statements of fact. Determine how the statement is related to the hypothesis and mark each statement according to the key which follows the hypothesis.

## GENERAL PROBLEM:

What factors are involved in the transmission and development of Infantile Paralysis (Poliomyelitis)?

## HYPOTHESIS I:

In man the disease is contracted by direct contact with persons having the disease.

For Items 52 through 60 mark space if the item offers:

1. Direct evidence in support of the hypothesis.
  2. Indirect evidence in support of hypothesis.
  3. Evidence which has no bearing on the hypothesis.
  4. Indirect evidence against the hypothesis.
  5. Direct evidence against the hypothesis.
52. Monkeys free from the disease almost never catch infantile paralysis from infected monkeys.
  53. The curve of number of cases of the disease in a given area is the same shape as the curve for the fly population in that area, the infantile paralysis incidence curve lagging behind the fly population curve by about two weeks.
  54. The virus has never been isolated from the blood.
  55. The virus is not found in the nasal secretion, nor in the saliva.
  56. The incubation period for infantile paralysis is from 4 to 21 days.
  57. Most persons in contact with the diseased individual do not develop the disease.
  58. The incidence of infantile paralysis is higher in rural districts than in the cities.
  59. Cases of infantile paralysis have been found to follow the roads of communication of the population, that is, the disease spreads from populated areas along roads or rivers to other areas.
  60. Even during epidemics cases are spotty, it is usually impossible to trace one case from another.
  61. What is the status of hypothesis I ?
    1. It is true.
    2. It is probably true.
    3. The data are contradictory, so the truth or falsity cannot be judged.
    4. The hypothesis is probably false.
    5. It is definitely false.

## HYPOTHESIS II:

Healthy persons having had contact with diseased individuals may carry the disease from one person to another.

For items 62 through 70 mark space if the item offers:

1. Direct evidence in support of the hypothesis.
  2. Indirect evidence in support of the hypothesis.
  3. Evidence which has no bearing on the hypothesis.
  4. Indirect evidence against the hypothesis.
  5. Direct evidence against the hypothesis.
- 
62. Monkeys free of the disease almost never catch infantile paralysis from infected monkeys.
  63. It has been found that exertion prior to or at the time of infection increases the incidence of the disease.
  64. Even during epidemics cases are spotty; it is usually impossible to trace one case from another.
  65. The virus is always found in the stools of people who have the disease.
  66. Most persons in contact with the diseased individual do not develop the disease.
  67. Nine out of 14 adults contacts had virus in stools, almost all child contacts have virus in stools.
  68. Up to two months after contact the virus is found in the stools of persons who contacted the victims, but who did not contract the disease.
  69. In the stools of non-contacts the virus was found in only one person in 100.
  70. The percent of cases of infantile paralysis is higher in rural districts than in the cities.
  71. What is the status of hypothesis II ?
    1. The hypothesis is true.
    2. It is probably false.
    3. The data are contradictory, so the truth or falsity cannot be judged.
    4. It is probably false.
    5. It is definitely false.

This portion of the test was designed to measure your ability to interpret data and to test your understanding of experimentation. In each case the numbers in the first column are the numbers which you will use as your answer. Thus the table presented becomes both the source of data and your key for the questions which follow it. In each case where a test tube number or group number is called for the one which gives positive evidence for the statement should be given. Below this the control or comparison is called for. This is the test tube or group number of the data which offers a comparison. For example:

1. Leaf in dark - no starch.
2. Leaf in light - starch.

"Light is necessary for the production of starch." You would mark space 2 because this is the positive evidence, but it would be meaningless if it were not compared with the leaf in the dark. Therefore, the following item, "What is the control (comparison) for item 1?" would be marked space 1.

Items 72 through 80 refer to the data presented below. Five test tubes, each containing a gram of protein, were set up. Mark each item according to the test tube number called for. All substances were dissolved in water. All test tubes were kept at 37° C. (water boils at 100° C.). For test tube 5, Substance X was boiled and then cooled before it was added to the protein.

<u>Test Tube</u>	<u>Contents of Tubes</u>	<u>Amt. of Substance W present after 24 hours.</u>
1	Protein plus Substance X	.05 gram
2	Protein plus water	.00 gram
3	Protein plus Substance X hydrochloric acid	.08 gram
4	Protein plus Hydrochloric acid	.00 gram
5.	Protein plus Substance X (boiled)	.00 gram
72.	Give the number of the test tube which acts as a control (comparison) for the entire experiment.	
73.	Give the number of the test tube which gives evidence that protein does not break down spontaneously into Substance W.	
74.	Give the number of the test tube which gives evidence that Substance X is the active substance in the breakdown of proteins.	



75. Give the number of the tube which is the control for item 74.
76. Give the number of the test tube which shows that a temperature of 37 degrees C. does not cause protein to break down into Substance W..
77. Which test tube gives evidence that Substance X is not a stable substance?
78. Which tube is the control for item 77.
79. Give the number of the test tube which indicates that hydrochloric acid alone is ineffective in breaking down proteins.
80. Give the control for item 79.

Items 81 through 91 refer to the data presented below. Mark each item according to the leaf number called for. Plant A normally stores starch in its leaves while Plant B does not normally store starch in its leaves.

The following experiments were performed in a dark room at 72° F. Glucose (sugar) solutions were made with 20 grams of glucose per 100 cubic centimeters of water. Leaves of plant A taken from a plant that had been in the dark for 48 hours were floated in the 5 solutions listed below and left in the glucose solution for an hour.

<u>Leaf</u>	<u>Solution</u>	<u>Analysis of leaf after 4 hours.</u>
1	Glucose	Starch in leaf
2	Water	No starch in leaf
3	Glucose plus juice from Plant B	No starch in leaf
4	Glucose plus juice from Plant C	No starch in leaf
5	Glucose plus boiled juice from Plant B	Small amount of starch in leaf

81. Give the number of the leaf which showed that starch does not develop spontaneously in the leaf in the dark.
82. This leaf indicates that a temperature of 72° F. does not cause starch to form in the leaf..
83. Give the number of the leaf which is the control (comparison) for the entire experiment.
84. Give the number of the leaf which gives evidence that Plant A is capable of manufacturing starch from glucose.

85. Give the number of the leaf which is the control for item 84.
86. Give the number of the leaf which gives evidence that the juice of Plant B is capable of preventing the manufacture of starch from glucose.
87. What is the control for item 86?
88. Give the number of the leaf which gives evidence that the juices of Plant B contain a substance which inhibits the production of starch in its leaves.
89. Give the leaf which is the control for item 88.
90. This leaf gives evidence that the inhibitory substance is not a stable substance.
91. What is the control for item 90?

This portion of the test was designed to measure your ability to make conclusions. When facts are analyzed and studied they sometimes yield evidence which help in the solution of a problem. However, any conclusion must be checked before it can be accepted. The following key includes four ways in which conclusions may be faulty. Each of the items present a question or problem, a brief description of an experiment and one or more conclusions drawn from the experiment. Each experiment was repeated many times. Read each problem, experiment and the conclusions. Where several conclusions are given evaluate each conclusion separately. Is the conclusion tentatively justified by the data? If so, mark space 1 on your answer sheet. If the conclusion is not justified determine whether 2, 3, 4, or 5 in the key is the best reason for it being faulty and mark the proper space on your answer sheet..

#### Key

The conclusion is:

1. Tentatively justified.
2. Unjustified because it does not answer the problem.
3. Unjustified because the experiment lacks a control comparison.
4. Unjustified because the data are faulty or inadequate, though a control was included.
5. Unjustified because it is contradicted by the data.

PROBLEM: A student was interested in developing a test for a certain type of substance. In all 100 cases his test was positive.

92. He concluded that the test was a specific test for the substance.

PROBLEM: An investigator wanted to know what causes people to breathe faster when they are running rapidly. He found that breathing more carbon dioxide increased the breathing rate, but that the breathing of air deficient in oxygen did not increase the breathing rate.

93. He concluded that people breathe faster when they are running because they need more oxygen.
94. Someone else concluded that running increases the rate of breathing.

PROBLEM: An investigator wished to determine whether temperature increased the rate of a certain reaction. On repeated tests he found that if he started out with a certain amount of his original substances he would obtain, after one hour, 1 gram of the substance produced by the reaction at  $0^{\circ}\text{C}$ ., 2 grams at  $20^{\circ}\text{C}$ ., 5 grams at  $40^{\circ}\text{C}$ ., and 3 grams at  $60^{\circ}\text{C}$ .

95. He concluded that increased temperature increased the rate of the reaction.

PROBLEM: A person wanted to determine whether bile aided in the digestion of fats. He found that whenever he mixed pancreatic juice with fats a small part of the fat was digested, but whenever he mixed pancreatic juice and bile with fat, he found that the fat was completely digested. When he mixed bile alone with fat he found that there was no digestion.

96. He concluded that bile aided in the digestion of fats.
97. Another concluded that pancreatic juice was necessary for digestion of fats.
98. Someone else claimed that bile does not aid in the digestion of fat.

PROBLEM: A person wanted to know what caused a certain disease. He examined 1000 patients with the disease. All had a certain bacteria (Bacteria A) in the digestive tract.

99. He concluded that Bacteria A was the cause of the disease.

PROBLEM: A person wanted to know why plants bend toward the light. He placed one group of plants in the light with the light source at the right. He placed another group of similar plants in the dark. The plants in the dark grew straight, the plants in the light were bent to the right.

100. He concluded that plants bend toward the light.
101. Another concluded that plants bend toward the light because they need light to grow.
102. Someone else concluded that light influences the direction in which plants grow.

PROBLEM: Investigator A wanted to know what caused people to become ill if confined in large numbers to a small closed area. He found on repeated tests that the air in very crowded closed areas contained about 5% carbon dioxide, while normal air contains .03% carbon dioxide.

103. He concluded that excessive carbon dioxide caused the illness.
104. Another investigator concluded that the illness was caused by insufficient oxygen.

PROBLEM: Investigator B in an attempt to solve the same problem repeated the experiment done by investigator A but in addition had people in uncrowded rooms breathe air containing 5% carbon dioxide. No ill effects were noted among those in the uncrowded rooms.

105. He also concluded that excessive carbon dioxide caused the illness.
106. Another investigator claimed that this showed that the disease was caused by insufficient oxygen.
107. Another conclusion was that 5% carbon dioxide will produce no ill effects.
108. Still another claimed that people live better in uncrowded areas.

PROBLEM: What are some of the requirements for seeds to sprout? The same student planted two groups of seeds of different types in pots and placed one group of the pots in the light, the others in the dark. Those plants in the light were green, those in the dark were yellow. Other conditions were the same for both groups.

109. Conclusion:

Light is necessary for sprouting of seeds.

110. Another conclusion:

Plants require light to mature properly.

This portion of the test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

### Key

1. True:

The data alone are sufficient to show that the statement is true.

2. Probably true:

The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.

3. Insufficient evidence:

There are no data to indicate whether there is any degree of truth or falsity in the statement.

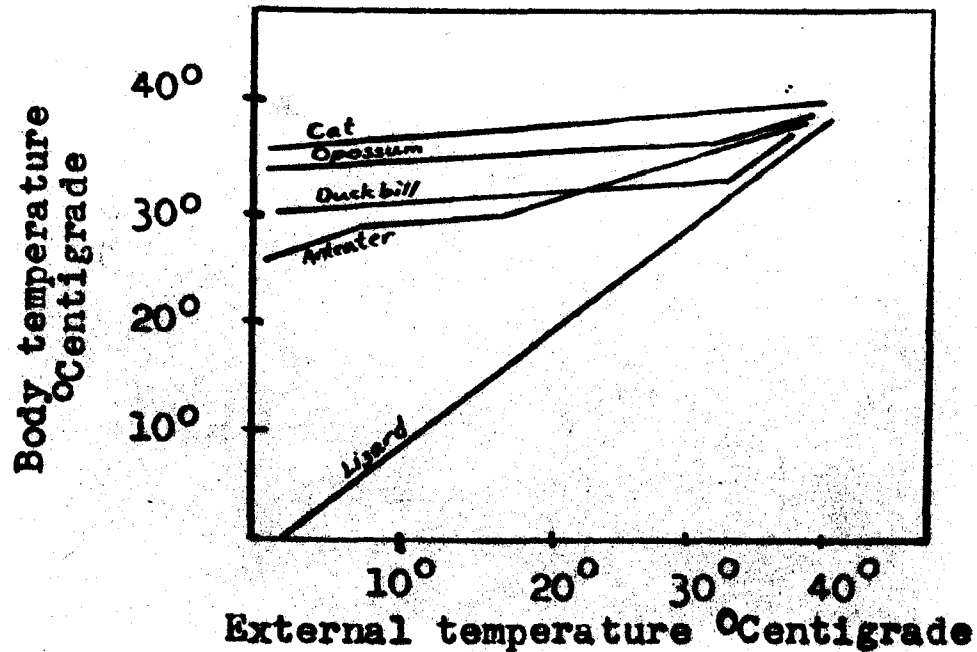
4. Probably false:

The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.

5. False:

The data alone are sufficient to show that the statement is false.

Items 111 through 131 refer to the following graph. Use the key above to answer the items. The lizard is considered to be cold blooded, the others warm blooded.



111. The body temperature of the cat varies more than the body temperature of the ant eater.
112. When the external temperature is 50°C., the temperature of the lizard is also 50°C.
113. At an external temperature of 50°C., the temperature of the cat is 50°C.
114. When the external temperature is 50°C., the temperature of the ant eater would be higher than the temperature of the cat.
115. The temperature of a mouse would be about half way between that of the cat and the ant eater.
116. At no time during the experiment did any of the animals have the same body temperature.
117. There is a close correlation between the body temperature of the lizard and that of the external environment.
118. The body temperature of the cat showed the least variation in temperature during the experimental period.
119. At 20 degrees below 0°C., the lizard would be frozen.
120. If the temperature of other cold blooded animals were plotted it would resemble that of the lizard.

Items 121 through 124 are a re-evaluation of some of the items 111 through 120. Re-read items 112, 114, 118 and 120 and determine whether they are generalizations, extensions of the data, explanations of the data or merely re-statements of the data, etc. Answer each according to the following key:

Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of cause and effect.
4. A restatement of results.
5. None of the above.

121. Item 112
122. Item 114
123. Item 118
124. Item 120

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data) The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 111 through 120.

Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: Warmblooded animals have some type of heat regulating mechanism.

- 125. It is possible for animals to have some type of heat regulating mechanism.
- 126. The cat and the duckbill are very different in their reaction to the external environment.
- 127. The opossum had a lower body temperature than the cat.

Conclusion II:

Ant eaters and duckbills are more closely related than ant eaters and cats.

- 128. Similarity of reaction of living things indicate a relationship.
- 129. The temperature of the ant eater varied more with the external temperature than did that of the cat.
- 130. The degree of closeness of similarity of response of living things runs parallel with the closeness of kinship.
- 131. The temperature of the cat varied less than that of the anteater and duckbill with change of temperature.

This portion of the test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

Key

- 1. True: The data alone are sufficient to show that the statement is true.
- 2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
- 3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
- 4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
- 5. False: The data alone are sufficient to show that the statement is false.



Analyses were made of the Vitamin C content of red ripe and green tomatoes as soon as they were picked. Mature green tomatoes were stored at the temperatures indicated in the following table. Those which had ripened by the end of the first week were analyzed for their Vitamin C content; those ripened at the end of the second week were analyzed at the end of the second week, etc. In addition some mature green tomatoes were analyzed each week.

<u>Condition when taken from field</u>	<u>Temp. when stored</u>	<u>No. of weeks stored</u>	<u>Stage of ripeness when analyzed</u>	<u>Vitamin C mg/100 grams</u>
mature green	not stored	0	mature green	15.0
red ripe	not stored	0	red ripe	16.2
mature green	70°F.	1	red ripe	14.4
mature green	70°F.	2	red ripe	12.9
mature green	70°F.	3	red ripe	8.2
mature green	80°F.	1	red ripe	14.0
mature green	80°F.	2	red ripe	9.8
mature green	80°F.	3	red ripe	7.1
mature green	70°F.	1	mature green	10.0
mature green	70°F.	2	mature green	7.2

132. Tomatoes ripened at 90°C. would have less Vitamin C after three weeks than those stored at 80°F.
133. Tomatoes could not be stored at 90°F. because at this high a temperature they would rot or spoil.
134. The lower the temperature at which tomatoes are stored the less is the breakdown of Vitamin C.
135. Heat causes a breakdown of the Vitamin C molecule.
136. After four weeks of storage tomatoes stored at 70°F., would contain less than 7 mg/100 grams of Vitamin C.
137. Some mature green tomatoes ripen in storage within a week.
138. The green tomatoes which did not ripen in a week had lost about the same amount of Vitamin C as those which ripened during the week.
139. Vitamin C is a stable substance.
140. Vitamin C is manufactured some place else in the plant than in the fruit (tomato) and is stored in the fruit.

Items 141 through 144 are a re-evaluation of some of the items 132 - 140. Re-read items 133, 135, 136 and 137 and determine whether they are generalizations, extensions of the data, explanations of the data or merely re-statements of the data, etc. Each of these items is to be answered according to the following key:

Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of causes and effect.
4. A restatement of results.
5. None of the above.

141. Item 133

142. Item 135

143. Item 136

144. Item 137

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data). The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items will relate to the data presented for items 81 through 100.

Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I:

Sunlight causes an increase in the Vitamin C content of tomatoes as they ripen on the vine.

145. The tomatoes which were analyzed when green ripe would have contained more Vitamin C if they had been allowed to ripen on the vine.
146. The test used to measure the amount of Vitamin C accurately measures the amount.
147. The Vitamin C content of ripe tomatoes on the vine was higher than the Vitamin C content of the green ripe tomatoes on the vines.

Conclusion II:

Vitamin C breaks down spontaneously at room temperature.

148. Vitamin C reacts similarly in all plants in which it is found.
149. When the tomatoes were stored at room temperature the Vitamin C content decreased.
150. All vitamins react similarly to storage at room temperature.

TABLE XLV  
ITEM ANALYSIS DATA FOR TEST I

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
1	*92.8	86.4	.16			
	**91.0	83.0	.17	10	86	73
2	77.6	56.8	.24			
	72.0	46.0	.24	15	59	55
3	80.0	55.2	.28			
	75.0	44.4	.32	20	59	55
4	96.8	72.0	.48			
	96.0	65.0	.50	33	80	68
5	96.8	71.2	.50			
	96.0	64.0	.51	34	80	68
6	56.0	36.0	.21			
	45.0	20.0	.27	17	31	40
7	88.0	63.2	.33			
	85.0	54.0	.36	23	68	60
8	78.4	67.2	.14			
	73.0	59.0	.17	10	66	59
9	48.8	56.0	-.10			
	36.0	45.0	-.10	- 6	40	45
10	80.8	47.2	.37			
	76.0	34.0	.43	28	55	53
11	41.6	20.8	.24			
	27.0	1.0	.60	42	14	27
12	44.8	47.2	-.03			
	31.0	34.0	-.05	- 2	31	40
13	88.8	60.8	.37			
	86.0	51.0	.40	26	68	60

\* Method of Flanagan  
\*\* Method of Davis

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
14	66.4	34.4	.38			
	58.0	18.0	.43	28	38	44
15	78.4	57.6	.24			
	73.0	47.0	.27	17	59	55
16	76.0	46.4	.33			
	70.0	33.0	.36	23	51	51
17	72.0	53.6	.21			
	65.0	42.0	.23	14	53	52
18	61.6	37.6	.24			
	52.0	22.0	.32	20	37	43
19	44.8	21.6	.26			
	31.0	2.0	.55	37	16	29
20	18.4	23.2	-.05			
	0.0	4.0	-.23	-14	3	9
21	46.4	37.6	.10			
	33.0	22.0	.14	8	27	37
22	60.0	28.0	.33			
	50.0	10.0	.48	32	30	39
23	69.6	56.8	.14			
	62.0	46.0	.17	10	53	52
24	71.2	58.4	.14			
	64.0	48.0	.17	10	53	52
25	68.8	40.0	.30			
	61.0	25.0	.36	23	24	46
26	67.2	45.6	.22			
	59.0	32.0	.27	17	25	47
27	80.0	52.8	.31			
	75.0	41.0	.35	22	57	54
28	50.4	33.6	.16			
	38.0	17.0	.26	16	27	37
29	53.6	24.8	.33			
	42.0	6.0	.50	33	24	35

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
30	62.4	44.8	.17			
	53.0	31.0	.23	14	40	45
31	53.6	27.2	.28			
	42.0	9.0	.41	27	25	36
32	55.2	27.2	.29			
	44.0	9.0	.45	29	25	36
33	60.8	45.6	.16			
	51.0	32.0	.20	12	40	45
34	36.8	10.4	.37			
	21.0	0.0	.55	37	12	25
35	20.8	11.2	.17			
	1.0	.0	.02	1	1	1
36	52.8	20.0	.36			
	41.0	0.0	.68	50	21	33
37	88.0	60.0	.36			
	85.0	50.0	.40	26	66	59
38	80.8	39.2	.44			
	76.0	24.0	.52	35	50	50
39	84.8	48.8	.41			
	81.0	36.0	.48	32	59	55
40	60.0	31.2	.30			
	50.0	14.0	.41	27	31	40
41	49.6	47.2	.03			
	37.0	34.0	.04	2	35	42
42	76.0	61.6	.16			
	70.0	52.0	.18	11	61	56
43	60.0	35.2	.26			
	50.0	19.0	.35	22	35	42
44	48.0	16.0	.38			
	35.0	0.0	.64	46	17	30
45	44.8	41.6	.04			
	31.0	27.0	.05	3	28	38

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
46	14.4 0.0	16.0 0.0	-.03 .00	0	0	0
47	84.4 81.0	49.6 37.0	.40 .46	30	57	54
48	72.0 62.0	41.6 25.0	.31 .38	24	42	46
49	80.8 75.0	57.6 46.0	.27 .32	20	61	56
50	60.8 51.0	43.2 29.0	.18 .22	13	40	45
51	76.8 69.0	64.0 55.0	.15 .15	9	61	56
52	42.4 28.0	29.6 12.0	.14 .23	14	19	32
53	52.0 40.0	16.8 0.0	.39 .68	50	21	33
54	74.4 68.0	47.2 34.0	.28 .35	22	51	51
55	48.8 36.0	28.0 10.0	.23 .36	23	22	34
56	89.6 87.0	68.8 61.0	.32 .32	20	73	63
57	90.4 88.0	76.0 70.0	.23 .26	16	79	67
58	61.6 51.0	40.0 27.0	.22 .26	16	38	44
59	56.8 46.0	47.2 34.0	.11 .14	8	40	45
60	56.0 45.0	32.8 16.0	.24 .34	21	30	39

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
61	62.4	22.4	.42			
	53.0	3.0	.66	48	28	38
62	64.0	40.8	.24			
	55.0	26.0	.31	19	40	45
63	89.6	60.8	.39			
	87.0	51.0	.41	27	68	60
64	55.2	21.6	.36			
	44.0	7.0	.50	33	25	36
65	69.6	40.8	.32			
	62.0	26.0	.36	23	44	47
66	32.0	18.4	.18			
	15.0	0.0	.47	31	8	20
67	26.4	21.6	.05			
	8.0	2.0	.24	15	5	15
68	28.0	26.4	.03			
	10.0	8.0	.05	3	9	22
69	52.8	27.2	.27			
	41.0	9.0	.41	27	24	35
70	48.8	39.2	.11			
	36.0	24.0	.14	8	30	39
71	61.2	47.2	.14			
	52.0	34.0	.18	11	42	46
72	81.6	54.4	.32			
	77.0	43.0	.36	23	59	55
73	70.4	21.6	.48			
	63.0	2.0	.73	56	31	40
74	71.2	41.6	.32			
	64.0	27.0	.38	24	44	47
75	34.4	15.2	.25			
	18.0	0.0	.51	34	9	22



TABLE XLV. (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
76	80.0	38.4	.44			
	75.0	23.0	.52	35	48	49
77	75.2	32.0	.43			
	69.0	15.0	.55	37	42	46
78	75.2	34.4	.42			
	69.0	18.0	.52	35	44	47
79	96.8	71.2	.55			
	96.0	64.0	.51	34	80	68
80	88.8	60.0	.37			
	86.0	50.0	.45	29	68	60
81	92.0	42.2	.58			
	90.0	28.0	.64	46	59	55
82	84.0	34.4	.52			
	80.0	18.0	.61	43	48	49
83	68.8	34.4	.36			
	61.0	18.0	.45	29	38	44
84	96.0	55.2	.58			
	95.0	44.0	.62	44	68	60
85	84.0	40.8	.47			
	80.0	26.0	.54	36	53	52
86	92.8	68.8	.39			
	91.0	61.0	.40	26	76	65
87	51.2	39.0	.29			
	39.0	5.0	.50	33	21	33
88	64.8	38.4	.27			
	56.0	23.0	.35	22	38	44
89	32.8	10.4	.34			
	16.0	0.0	.48	32	9	21
90	94.4	48.0	.58			
	93.0	35.0	.64	46	63	57

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
91	84.8	40.8	.53			
	81.0	26.0	.55	37	53	52
92	60.8	32.0	.31			
	51.0	15.0	.41	27	31	40
93	76.8	40.0	.38			
	71.0	25.0	.46	30	46	48
94	72.8	45.6	.28			
	66.0	32.0	.35	22	48	49
95	52.8	21.6	.34			
	41.0	2.0	.62	44	21	33
96	91.2	56.8	.45			
	89.0	46.0	.48	32	66	59
97	25.6	5.6	.38			
	7.0	0.0	.32	20	4	13
98	90.4	60.4	.40			
	88.0	51.0	.47	31	70	61
99	88.0	49.6	.45			
	85.0	37.0	.51	34	59	55
100	73.6	33.6	.41			
	67.0	11.0	.59	41	38	44
101	51.2	23.2	.30			
	39.0	4.0	.54	36	21	33
102	49.6	18.4	.36			
	37.0	0.0	.66	48	18	31
103	20.0	6.4	.28			
	0.0	0.0	.00	0	0	0
104	48.0	18.4	.35			
	35.0	0.0	.65	47	17	30
105	85.6	55.2	.37			
	82.0	44.0	.40	26	63	57

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
106	52.8 41.0	22.4 3.0	.34 .58	40	21	33
107	28.0 10.0	14.4 0.0	.20 .39	25	5	16
108	72.0 65.0	44.0 30.0	.28 .35	22	46	48
109	66.4 58.0	32.8 16.0	.34 .45	29	37	43
110	32.0 15.0	17.6 0.0	.20 .47	31	8	20
111	85.6 82.0	69.6 62.0	.23 .24	15	71	62
112	66.4 58.0	26.4 8.0	.41 .58	40	33	41
113	69.6 62.0	28.0 10.0	.43 .57	39	35	42
114	81.6 77.0	37.6 22.0	.47 .55	37	48	49
115	94.4 93.0	78.4 73.0	.31 .34	21	82	69
116	86.4 83.0	56.0 45.0	.37 .40	26	63	57
117	93.6 92.0	65.6 57.0	.44 .46	30	74	64
118	63.2 54.0	55.2 44.0	.08 .10	6	48	49
119	68.8 61.0	35.2 19.0	.34 .45	29	38	44
120	50.4 38.0	33.6 17.0	.17 .26	16	27	37

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
121	75.2 69.0	48.8 36.0	.28 .34	21	51	51
122	84.0 80.0	60.8 51.0	.28 .32	20	64	58
123	61.6 52.0	38.4 23.0	.24 .31	19	38	44
124	79.2 74.0	44.2 31.0	.36 .41	27	53	52
125	76.0 70.0	60.8 51.0	.17 .20	12	61	56
126	23.2 4.0	20.0 0.0	.04 .23	14	3	9
127	65.6 57.0	34.4 18.0	.35 .41	27	37	43
128	58.4 48.0	44.8 31.0	.13 .17	10	40	45
129	70.4 63.0	28.0 10.0	.42 .57	39	35	42
130	49.6 37.0	24.8 6.0	.27 .46	30	21	33
131	68.0 60.0	36.0 20.0	.33 .41	27	40	45
132	59.2 49.0	32.0 15.0	.29 .39	25	31	40
133	92.0 90.0	55.2 44.0	.48 .48	32	66	59
134	32.8 16.0	24.0 5.0	.11 .26	16	10	23
135	48.0 35.0	29.6 12.0	.19 .31	19	22	34

TABLE XLV (continued)

Item	Percent Success		Discrimination		Difficulty	
	Upper 27%	Lower 27%	r	Index	% Success	Index
136	72.0	24.8	.47			
	65.0	6.0	.65	47	35	42
137	89.6	66.4	.34			
	87.0	58.0	.35	22	71	62
138	72.0	32.0	.41			
	65.0	15.0	.52	35	38	44
139	24.0	12.8	.17			
	5.0	0.0	.26	16	3	10
140	60.8	54.4	.07			
	51.0	43.0	.08	5	46	48
141	59.2	21.6	.38			
	49.0	2.0	.67	49	25	36
142	44.8	25.6	.21			
	31.0	7.0	.38	24	18	31
143	83.2	28.8	.55			
	79.0	11.0	.68	50	44	47
144	81.6	37.6	.47			
	77.0	22.0	.55	37	48	49
145	76.8	50.4	.30			
	71.0	38.0	.34	21	53	52
146	40.0	17.6	.28			
	25.0	0.0	.56	38	12	25
147	51.2	21.6	.32			
	39.0	2.0	.61	43	19	32
148	24.0	19.2	.06			
	5.0	0.0	.26	16	3	10
149	41.6	21.6	.23			
	27.0	2.0	.52	35	14	27
150	*69.6	41.6	.29			
	**62.0	27.0	.35	22	44	47

\* Method of Flanagan

\*\* Method of Davis

### APPENDIX III

## TEST IA

## THE ABILITY TO THINK SCIENTIFICALLY

## GENERAL DIRECTIONS

1. Place your name, age and sex in the spaces provided on the answer sheet.
2. Place your student number in the space provided for "date of birth".
3. On the space marked "school" place your major.
4. In the space marked "1" below "school", give courses you have had in science in high school, in the space marked "2" give any courses you have had in science in college in addition to biological science.
5. Answer all items; if you don't know - guess.
6. Do not mark on the test booklet. Use scratch paper if you wish.
7. Be sure to mark dark on the answer sheet; the machine does not pick up light markings.
8. Each item has only one answer; do not mark more than one.

This test has been devised to measure your ability to think scientifically. It is divided into several parts, each of these parts tests a different phase of scientific thinking.

This portion of the test is designed to measure your ability to differentiate phases of thinking. These steps include major problems or perplexities, possible solutions to problems, observations which are not results of experimentation but rather preliminary observations, results of experimentation, and conclusions.

The following key is to be used for the succeeding paragraph. Certain parts of the paragraph are underlined, and each underlined item is a question. Choose the proper response from the key and blacken the appropriate space in the answer sheet.

# KEY

1. A major problem (stated or implied).
2. Hypothesis (possible solution to problem).
3. Result of experimentation.
4. Initial observation (not experimental).
5. Conclusion (probable solution of problem).

(1) How does a homing pigeon navigate over territory it has never seen before? (2) Do air currents stimulate the pigeon in some way? (3) Are the pigeons equipped with some sort of magnetic compasses; that is, are they sensitive to the earth's magnetism? Yeagley tested the latter by fastening small magnets to the wings of well-trained pigeons. (4) Most of these birds never got home. (5) Others, carrying equal wing weights of non-magnetic copper, made the home roost without trouble, (6) indicating that the earth's magnetism is a factor in pigeon navigation. But the pigeons magnetic compass could not, by itself, bring him back to his roost; because many places on the earth's surface have identical magnetic conditions. Yeagley endeavored (7) to determine the other guiding factor. (8) It might be the sun or



stars, but pigeons navigate under clouds. While looking at a map which had lines representing the intensity of the earth's magnetism,

# Abbreviated Key

1. A major problem
2. Hypothesis
3. Results
4. Observations
5. Conclusions

he noted that the lines were crossed at varying angles by the parallels of latitude. If pigeons are sensitive to some factor connected with the lines of latitude, they would have all they need to find their way home. The next step was (9) to find some physical force, something the pigeons might be able to detect, related to the lines of latitude. The effect of the earth's turning varies directly with latitude; objects near the equator are carried daily around the earth's circumference, moving at over 1,000 mi. per hr. Objects near the poles are carried around more slowly. The direction and variation of this circling can be recorded by various man-made instruments. (10) Why shouldn't the pigeons feel it, too? If they could, they would have, along with their magnetic compass, a satisfactory navigating instrument. Yeagley trained hundreds of pigeons to return to their home roosts at State College, Pa. Then he took them to a part of Nebraska where the lines representing the earth's magnetism cross the parallels of latitude at the same angle as at State College. He released the pigeons to the east of this spot. (11) The pigeons all flew west. Yeagley believes that (12) pigeons are guided by both the earth's magnetism and by its turning. (13) Just where the birds keep their instruments is still unknown; but Yeagley

found that (14) birds have a mysterious organ in their eyes at the end of the optic nerve. (15) This organ may contain the nerve fibers that pick up vibrations of magnetism and the even more delicate sense that measure the earth's turning.

This portion of the test is designed to measure your ability to recognize faulty experimental procedures. In each case a problem and a possible solution to the problem (an hypothesis) are presented. In each case the experiments were designed by students to test the hypotheses. Judge each experiment according to the following key.

#### Key

This experiment is:

1. Satisfactory.
2. Unsatisfactory because it lacks a control or comparison.
3. Unsatisfactory because the control or comparison is faulty.
4. Unsatisfactory because it is unrelated to the hypothesis.
5. None of the above - the experiment is unsatisfactory for reasons other than those listed in 2, 3, and 4.

PROBLEM: What are some of the requirements for the sprouting of seeds?

HYPOTHESIS:

Oxygen is a requirement for the sprouting of seeds.

16. If a seed lacked oxygen under a controlled experiment the seed would not function properly and would soon die.
17. Place growing plants in an air tight container. Pump out the oxygen. Place other growing plants in containers with oxygen. Keep temperature, light, etc., the same for each.
18. Plant seeds in a container with glass covering it so that no oxygen can enter and see if they sprout. Keep temperature, light and moisture normal.

**PROBLEM:** A minute insect (aphid) is suspected of spreading a virus disease of roses. How would you determine whether this is true?

**HYPOTHESIS:** The aphid spreads a virus disease of roses.

19. Put the insect among other kinds of plants other than roses. Leave another group of these plants free from contact with the aphids. Compare the results.
20. Since aphids travel through the air, a plot of roses must be entirely protected from them, and another exposed to aphids which in turn have been exposed to roses afflicted with the virus disease. All must be under constant conditions of soil, atmosphere, etc.
21. Take sample rose with the virus disease. Obtain same kind of rose with no disease. Use microscope to aid in detection of the disease. Use some sort of spray. Note results.
22. In order to determine whether the aphid spreads a virus disease in roses, a group of roses should be put in a hot house free from aphids to see whether they get such a virus disease.

**PROBLEM:** To determine the cause of illness which appears when large numbers of people are confined to a small space.

**HYPOTHESIS:** Lack of oxygen causes the people to become ill.

23. One might check the oxygen by placing a number of people in a confined place where there was a control amount. Other checks would have to be made also such as the purity of food, the purity of water and whether or not proper sanitation rules were followed.
24. Put one group of people in a room with an excessive amount of carbon dioxide and another group in a room with a normal amount of carbon dioxide. Keep the oxygen concentration the same in both rooms.

This test is designed to measure your understanding of the relation of facts to the solution of a problem. The over-all problem involved in this test is presented. This is followed by a series of possible solutions to the problem (hypotheses). After each hypothesis there are a number of items, all of which are true statements of fact. Determine how the statement is related to the hypothesis and mark each statement according to the key which follows the hypothesis.

GENERAL PROBLEM: What factors are involved in the transmission and development of Infantile Paralysis (Poliomyelitis)?

HYPOTHESIS I: In man the disease is contracted by direct contact with persons having the disease.

For items 25 through 34 mark space if the item offers:

1. Direct evidence in support of the hypothesis.
2. Indirect evidence in support of the hypothesis.
3. Evidence which has no bearing on the hypothesis.
4. Indirect evidence against the hypothesis.
5. Direct evidence against the hypothesis.

25. Monkeys free from the disease almost never catch infantile paralysis from infected monkeys.
26. The curve of number of cases of the disease in a given area is the same shape as the curve for the fly population in that area, the infantile paralysis incidence curve lagging behind the fly population curve by about two weeks.
27. The virus has never been isolated from the blood.
28. The virus is not found in the nasal secretion nor in the saliva.
29. The incubation period for infantile paralysis is from 4 to 21 days.
30. Most persons in contact with the diseased individual do not develop the disease.
31. The incidence of infantile paralysis is higher in rural districts than in the cities.
32. Cases of infantile paralysis have been found to follow the roads of communication of the population, that is, the disease spreads from populated areas along roads or rivers to other areas.

33. Even during epidemics cases are spotty, it is usually impossible to trace one case from another.
34. What is the status of hypothesis I?
1. It is true.
  2. It is probably true.
  3. The data are contradictory, so the truth or falsity cannot be judged.
  4. The hypothesis is probably false.
  5. It is definitely false.

HYPOTHESIS II: Healthy persons having had contact with diseased individuals may carry the disease from one person to another.

For items 35 through 44 mark space if the item offers:

1. Direct evidence in support of the hypothesis.
2. Indirect evidence in support of the hypothesis.
3. Evidence which has no bearing on the hypothesis.
4. Indirect evidence against the hypothesis.
5. Direct evidence against the hypothesis.

35. Monkeys free of the disease almost never catch infantile paralysis from infected monkeys.
36. It has been found that exertion prior to or at the time of infection increases the incidence of the disease.
37. Even during epidemics cases are spotty; it is usually impossible to trace one case from another.
38. The virus is always found in the stools of people who have the disease.
39. Most persons in contact with the diseased individual do not develop the disease.
40. Nine out of 14 adult contacts had virus in stools, almost all child contacts have virus in stools.
41. Up to two months after contact the virus is found in the stools of persons who contacted the victims, but who did not contract the disease.
42. In the stools of non-contacts the virus was found in only one person in 100.
43. The percent of cases of infantile paralysis is higher in rural districts than in the cities.

44. What is the status of hypothesis II?

1. The hypothesis is true.
2. It is probably true.
3. The data are contradictory, so the truth or falsity cannot be judged.
4. It is probably false.
5. It is definitely false.

This portion of the test was designed to measure your ability to interpret data and to test your understanding of experimentation. In each case the numbers in the first column are the numbers which you will use as your answer. Thus the table presented becomes both the source of data and your key for the questions which follow it. In each case where a test tube number or group number is called for the one which gives positive evidence for the statement should be given. Below this the control or comparison is called for. This is the test tube or group number of the data which offers a comparison. For example:

1. Leaf in dark - no starch.
2. Leaf in light - starch.

"Light is necessary for the production of starch." You would mark space 2 because this is the positive evidence, but it would be meaningless if it were not compared with the leaf in the dark. Therefore, the following item, "What is the control (comparison) for item 1?" would be marked space 1.

Items 45 through 53 refer to the data presented below. Five test tubes, each containing a gram of protein, were set up. Mark each item according to the test tube number called for. All substances were dissolved in water. All test tubes were kept at 37° C. (water boils at 100° C.) For test tube 5, Substance X was boiled and then cooled before it was added to the protein.

<u>Test Tube</u>	<u>Contents of Tubes</u>	<u>Amt. of Substance W present after 24 hours</u>
1	Protein plus Substance X	.05 gram
2	Protein plus Water	.00 gram
3	Protein plus Substance X hydrochloric acid	.08 gram
4	Protein plus Hydrochloric acid	.00 gram
5	Protein plus Substance X (boiled)	.00 gram

45. Give the number of the test tube which acts as a control (comparison) for the entire experiment.
46. Give the number of the test tube which gives evidence that protein does not break down spontaneously into Substance W.
47. Give the number of the test tube which gives evidence that Substance X is the active substance in the break down of proteins.
48. Give the number of the tube which is the control for item 47.
49. Give the number of the test tube which shows that a temperature of 37° C. does not cause protein to break down into Substance W.
50. Which test tube gives evidence that Substance X is not a stable substance?
51. Which test tube is the control for item 50?
52. Give the number of the test tube which indicates that hydrochloric acid alone is ineffective in breaking down proteins.
53. Give the control for item 52.

Items 54 through 64 refer to the data presented below. Mark each item according to the leaf number called for. Plant A normally stores starch in its leaves while Plant B does not normally store starch in its leaves. The following experiments were performed in a dark room at 72° F. Glucose (sugar) solutions were made with 20 grams of glucose per 100 cubic centimeters of water. Leaves of plant A taken from a plant that had been in the dark for 48 hours were floated in the 5 solutions listed below and left in the glucose solution for an hour.

<u>Leaf</u>	<u>Solution</u>	<u>Analysis of leaf after 4 hours</u>
1	Glucose	Starch in leaf
2	Water	No starch in leaf
3	Glucose plus juice from Plant B	No starch in leaf
4	Glucose plus juice from Plant C	No starch in leaf
5	Glucose plus boiled juice from Plant B	Small amount of starch in leaf

54. Give the number of the leaf which showed that starch does not develop spontaneously in the leaf in the dark.
55. This leaf indicates that a temperature of 72° F. does not cause starch to form in the leaf.
56. Give the number of the leaf which is the control (comparison) for the entire experiment.
57. Give the number of the leaf which gives evidence that Plant A is capable of manufacturing starch from glucose.
58. Give the number of the leaf which is the control for item 57.
59. Give the number of the leaf which gives evidence that the juice of Plant B is capable of preventing the manufacture of starch from glucose.
60. What is the control for item 59?
61. Give the number of the leaf which gives evidence that the juices of Plant B contain a substance which inhibits the production of starch in its leaves.
62. Give the leaf which is the control for item 61.
63. This leaf gives evidence that the inhibitory substance is not a stable substance.
64. Give the control for item 63.

This portion of the test was designed to measure your ability to make conclusions. When facts are analyzed and studied they sometimes yield evidence which help in the solution of a problem. However, any conclusion must be checked before it can be accepted. The following key includes four ways in which conclusions may be faulty. Each of the items present a question or problem, a brief description of an experiment and one or more conclusions drawn from the experiment. Each experiment was repeated many times. Read each problem, experiment and the conclusions. Where several conclusions are given evaluate each conclusion separately. Is the conclusion tentatively justified by the data? If so, mark space 1 on your answer sheet. If the conclusion is not justified determine whether 2, 3, 4, or 5 in the key is the best reason for it being faulty and mark the proper space on your answer sheet.



Key

The conclusion is:

1. Tentatively justified.
2. Unjustified because it does not answer the problem.
3. Unjustified because the experiment lacks a control comparison.
4. Unjustified because the data are faulty or inadequate, though a control was included.
5. Unjustified because it is contradicted by the data.

PROBLEM: A student was interested in developing a test for a certain type of substance. In all 100 cases his test was positive.

65. He concluded that the test was a specific test for the substance.

PROBLEM: An investigator wanted to know what causes people to breathe faster when they are running rapidly. He found that breathing more carbon dioxide increased the breathing rate, but that the breathing of air deficient in oxygen did not increase the breathing rate.

66. He concluded that people breathe faster when they are running because they need more oxygen.
67. Someone else concluded that running increases the rate of breathing.

PROBLEM: An investigator wished to determine whether temperature increased the rate of a certain reaction. On repeated tests he found that if he started out with a certain amount of his original substances he would obtain, after one hour, 1 gram of the substance produced by the reaction at 0°C., 2 grams at 20°C., 5 grams at 40°C. and 3 grams at 60°C.

68. He concluded that increased temperature increased the rate of the reaction.

PROBLEM: A person wanted to determine whether bile aided in the digestion of fats. He found that whenever he mixed pancreatic juice with fats a small part of the fat was digested, but whenever he mixed pancreatic juice and bile with fat, he found that the fat was completely digested. When he mixed bile alone with fat he found that there was no digestion.

69. He concluded that bile aided in the digestion of fats.
70. Another concluded that pancreatic juice was necessary for digestion of fats.
71. Someone else claimed that bile does not aid in the digestion of fat.

PROBLEM: A person wanted to know what caused a certain disease. He examined 1000 patients with the disease. All had a certain bacteria (Bacteria A) in the digestive tract.

72. He concluded that Bacteria A was the cause of the disease.

PROBLEM: A person wanted to know why plants bend toward the light. He placed one group of plants in the light with the light source at the right. He placed another group of similar plants in the dark. The plants in the dark grew straight, the plants in the light were bent to the right.

73. He concluded that plants bend toward the light.
74. Another concluded that plants bend toward the light because they need light to grow.
75. Someone else concluded that light influences the direction in which plants grow.

PROBLEM: Investigator A wanted to know what caused people to become ill if confined in large numbers to a small closed area. He found on repeated tests that the air in very crowded closed areas contained about 5% carbon dioxide, while normal air contains .03% carbon dioxide.

76. He concluded that excessive carbon dioxide caused the illness.
77. Another investigator concluded that the illness was caused by insufficient oxygen.

PROBLEM: Investigator B in an attempt to solve the same problem repeated the experiment done by investigator A but in addition had people in uncrowded rooms breathe air containing 5% carbon dioxide. No ill effects were noted among those in the uncrowded rooms.

78. He also concluded that excessive carbon dioxide caused the illness.
79. Another investigator claimed that this showed that the disease was caused by insufficient oxygen.
80. Another conclusion was that 5% carbon dioxide will produce no ill effects.
81. Still another claimed that people live better in uncrowded areas.

PROBLEM: What are some of the requirements for seeds to sprout? The same student planted two groups of seeds of different types in pots and placed one group of the pots in the light, the others in the dark. Those plants in the light were green, those in the dark were yellow. Other conditions were the same for both groups.

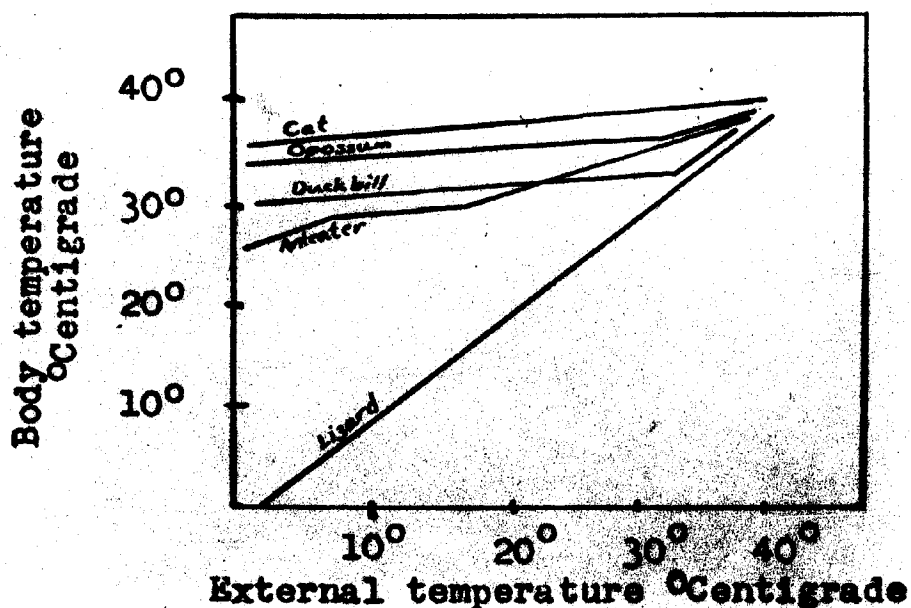
82. Conclusion: Light is necessary for sprouting of seeds.

This portion of the test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

#### Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

Items 83 through 102 refer to the following graph. Use the key above to answer the items. The lizard is considered to be cold blooded, the others warm blooded.



83. The body temperature of the cat varies more than the body temperature of the ant eater.
84. When the external temperature is 50° C., the temperature of the lizard is also 50° C.
85. At an external temperature of 50° C., the temperature of the cat is 50° C.
86. When the external temperature is 50° C., the temperature of the ant eater would be higher than the temperature of the cat.
87. The temperature of a mouse would be about half way between that of the cat and the ant eater.
88. At no time during the experiment did any of the animals have the same body temperature.
89. There is a close correlation between the body temperature of the lizard and that of the external environment.
90. The body temperature of the cat showed the least variation in temperature during the experimental period.
91. At 20 degrees below 0° C. the lizard would be frozen.
92. If the temperature of other cold blooded animals were plotted it would resemble that of the lizard.

Items 93 through 96 are a re-evaluation of some of the items 83 through 92. Re-read items 93, 94, 95 and 96 and determine whether they are generalizations, extensions of the data, explanations of the data or merely restatements of the data, etc. Answer each according to the following key:

Key

1. A generalization, that is the data says it is true for this situation; a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of cause and effect.
4. A restatement of results.
5. None of the above.

93. Item 84

94. Item 86

95. Item 90

96. Item 92

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data) The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items all relate to the data presented for items 83 through 92.

Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: Warmblooded animals have some type of heat regulating mechanism.

97. It is possible for animals to have some type of heat regulating mechanism.

98. The opossum had a lower body temperature than the cat.

Conclusion II: Ant eaters and duckbills are more closely related than ant eaters and cats..

99. Similarity of reaction of living things indicate a relationship.
100. The temperature of the ant eater varied more with the external temperature than did that of the cat.
101. The degree of closeness of similarity of response of living things runs parallel with the closeness of kinship.
102. The temperature of the cat varied less than that of the ant eater and duckbill with change of temperature.

This portion of the test was designed to measure your ability to interpret data. Following the data you will find a number of statements. You are to assume that the data as presented are true. Evaluate each statement according to the following key and mark the appropriate space on your answer sheet.

#### Key

1. True: The data alone are sufficient to show that the statement is true.
2. Probably true: The data indicate that the statement is probably true, that it is logical on the basis of the data but the data are not sufficient to say that it is definitely true.
3. Insufficient evidence: There are no data to indicate whether there is any degree of truth or falsity in the statement.
4. Probably false: The data indicate that the statement is probably false, that is, it is not logical on the basis of the data but the data are not sufficient to say that it is definitely false.
5. False: The data alone are sufficient to show that the statement is false.

Analyses were made of the Vitamin C content of red ripe and green tomatoes as soon as they were picked. Mature green tomatoes were stored at the temperatures indicated in the following table. Those which had ripened by the end of the first week were analyzed for their Vitamin C content; those ripened at the end of the second week were analyzed at the end of the second week, etc. In addition some mature green tomatoes were analyzed each week.

<u>Condition when taken from field</u>	<u>Temp. when stored</u>	<u>No. of weeks stored</u>	<u>Stage of ripeness when analyzed</u>	<u>Vitamin C mg/100 grams</u>
mature green	not stored	0	mature green	15.0
red ripe	not stored	0	red ripe	16.2
mature green	70°F.	1	red ripe	14.4
mature green	70°F.	2	red ripe	12.9
mature green	70°F.	3	red ripe	8.2
mature green	80°F.	1	red ripe	14.0
mature green	80°F.	2	red ripe	9.8
mature green	80°F.	3	red ripe	7.1
mature green	70°F.	1	mature green	10.0
mature green	70°F.	2	mature green	7.2

103. Tomatoes ripened at 90°C. would have less Vitamin C after three weeks than those stored at 80°F.
104. Tomatoes could not be stored at 90°F. because at this high a temperature they would rot or spoil.
105. The lower the temperature at which tomatoes are stored the less is the breakdown of Vitamin C.
106. Heat causes a breakdown of the Vitamin C molecule.
107. After four weeks of storage tomatoes stored at 70°F. would contain less than 7 mg/100 grams of Vitamin C.
108. Some mature green tomatoes ripen in storage within a week.
109. The green tomatoes which did not ripen in a week had lost about the same amount of Vitamin C as those which ripened during the week.
110. Vitamin C is a stable substance.
111. Vitamin C is manufactured some place else in the plant than in the fruit (tomato) and is stored in the fruit.

Items 112 through 115 are a re-evaluation of some of the items 103 - 111. Re-read items 105, 106, 107 and 108 and determine whether they are generalizations, extensions of the data, explanations of the data or merely restatements of the data, etc. Each of these items is to be answered according to the following key:

Key

1. A generalization, that is the data says it is true for this situation, a generalization says it is true for all similar situations.
2. The data indicates a trend which if continued in either direction would make the statement true.
3. An explanation of the data in terms of causes and effect.
4. A restatement of results.
5. None of the above.

112. Item 105

113. Item 106

114. Item 107

115. Item 108

This phase of the test is designed to measure your understanding of assumptions underlying conclusions. A conclusion is given. (This conclusion is not necessarily justified by the data) The statements which follow the conclusion are the items which are to be evaluated according to the following key. These items will relate to the data presented for items 103 through 111.

Key

1. An assumption which must be made to make the conclusion valid (true).
2. An assumption which if made would make the conclusion false.
3. An assumption which has no relation to the validity (truth) of the conclusion.
4. Not an assumption; a restatement of fact.
5. Not an assumption; a conclusion.

Conclusion I: Sunlight causes an increase in the Vitamin C content of tomatoes as they ripen on the vine.





116. The tomatoes which were analyzed when green ripe would have contained more Vitamin C if they had been allowed to ripen on the vine.
117. The test used to measure the amount of Vitamin C accurately measures the amount.
118. The Vitamin C content of ripe tomatoes on the vine was higher than the Vitamin C content of the green ripe tomatoes on the vines.

Conclusion II: Vitamin C breaks down spontaneously at room temperature.



119. Vitamin C reacts similarly in all plants in which it is found.
120. When the tomatoes were stored at room temperature the Vitamin C content decreased.
121. All vitamins react similarly to storage at room temperature.

This portion of the test is designed to test your ability to organize data. Select from the key below the curve which best fits the data. If none of the curves fit the data mark space five on your answer sheet. The curves need not have the same amount of slope as the curves presented in the key. Use scratch paper if you wish.

1. 
2. 
3. 
4. 
5. None of the curves.

122. The horizontal axis represents the time in hours after the injection of sugar into the blood; the vertical axis is the amount of sugar in the blood.

<u>Time after injection</u>	<u>Blood sugar</u>
1	35
3	12
6	8

123. The horizontal axis represents age in years. The vertical axis is the percent increase in the weight of the ovaries and other female sex organs from birth to 20 years.

<u>Age</u>	<u>Percent increase</u>
4	8
10	12
14	20
18	80

124. The horizontal axis represents time in days; the vertical axis is the number of yeast cells in millions (starting with 100 yeast cells).

<u>Time in days</u>	<u>Number of yeast cells in millions</u>
4	25
8	150
12	390
20	400

125. The horizontal axis represents the amount of thyroprotein fed daily to cows. The vertical axis represents the percent increase in milk production.

<u>Thyropotein fed</u>	<u>Percent increase</u>
.15 grams	18
.20 grams	23
.24 grams	27
.30 grams	33

## APPENDIX IV

## RATING SCALE FOR ABILITY TO USE SCIENTIFIC METHOD

Person Rated

Rater

Date

**Directions:** Will you please rate the person whose name appears above on the two following characteristics. The two extremes of these characteristics are described. Place a cross (X) on the line indicating your judgment of the individual with respect to the qualities in question.

### 1. Ability to evaluate and devise experiments

Very superior	Superior	Average	Inferior	Very inferior
High degree of ability: Includes control factors, controls all but one variable, understands problem and devises experiments to test hypothesis. Can devise experiments which will yield results, recognizes problems inherent in the experiment, and has an understanding of what is happening in the experiment.			Low degree of ability: Experiments lack control or control is faulty, experiment unrelated to hypothesis. Student does not understand the experimental set-up, or the problems inherent in the experiment.	

### 2. Ability to interpret data (ability to form hypotheses and draw conclusions)

Very superior	Superior	Average	Inferior	Very inferior
High degree of ability: Is able to make logical inferences from data, takes pertinent facts into consideration, applies previous knowledge to the new situation, is able to see relationships, especially cause and effect relationships. Knows what evidence for his inference is, and why it is evidence.			Low degree of ability: Is unable to make logical inferences from data, does not differentiate between relevant and irrelevant data or between critical and non-critical data, is unable to see relationships.	